BIG DATA

一本书读懂大数据

黄 颖◎编著

看懂大数据　掌握时代先机

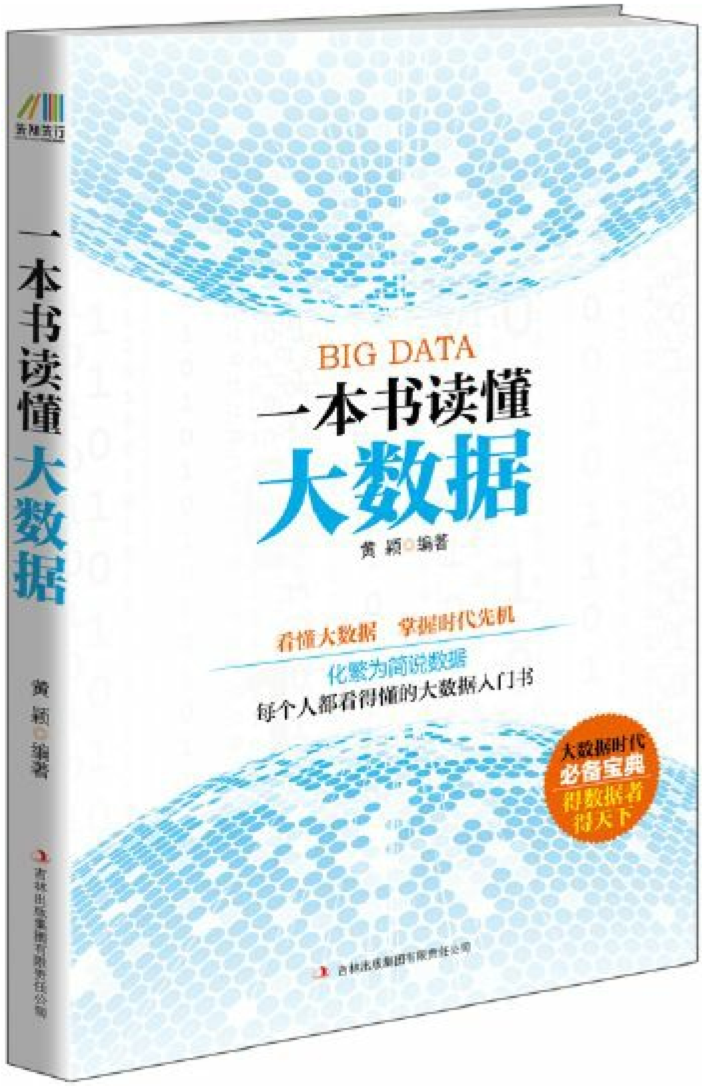
化繁为简说数据

每个人都看得懂的大数据入门书

大数据时代
必备宝典
得数据者
得天下

吉林出版集团有限责任公司

一本书读懂大数据
黄颖◎编著
吉林出版集团有限责任公司


BIG DATA
一本书读懂
大数据
黄 颖◎编著
看懂大数据 掌握时代先机
化繁为简说数据
每个人都看得懂的大数据入门书
大数据时代
必备宝典
得数据者
得天下
吉林出版集团有限责任公司

版权

图书在版编目（CIP）数据

一本书读懂大数据/黄颖编著.—长春：吉林出版集团有限责任公司，2014.11

ISBN 978-7-5534-5736-9

Ⅰ.一… Ⅱ.黄… Ⅲ.商业信息学 Ⅳ.F713.51

中国版本图书馆CIP数据核字（2014）第225942号

一本书读懂大数据

编者 黄颖

策划编辑 李异鸣

特约编辑 周乔蒙

责任编辑 齐琳 王平

封面设计 上尚装帧

出版 吉林出版集团有限责任公司

电话 总编办：010-63109269

发行部：010-81282844

前言

世界的万千变化一直超乎人们的预测，自2012年以来，大数据一词成了人类生活的代名词。如今，数据几乎已经渗透到了每一个行业的每一个领域之中，成了不可或缺的生产因素。每一天，互联网都会繁衍出无数的数据，这些内容足以刻满2亿张光碟；而手机客户端发出的帖子和邮件总数可达到3000万亿……如此惊人的数据使得对海量数据的挖掘和分析，成了企业发展的重要内容。大数据的数量大、类型多、时效快、价值密度低的特点，让这个崭新的时代充满了变数和乐趣。

数据迅速地膨胀，让差别细微的算法就足以决定企业的发展方向。很多企业在大数据时代纷纷进行了多种多样的尝试，这是一场伟大的革命，庞大的数据资源的冲击，让商界、学术界等所有领域都开始了量化的流程，积极探索大数据时代的奥秘。

这本书中，我们能够揭开大数据的面纱，挖掘和分析大数据整个流程的重要关卡，掌握大数据的多种特性和价值特征，对其结构有精准的把握。同时，我们将大数据和小数据时代进行对比，让读者更加清晰地认识我们生活的时代。

电子商务界乘着大数据的狂风，掀起了人们生活的数据风暴。国内外的企业使尽浑身解数，在大数据的海洋里摸爬滚打。从市场来看，阿里巴巴、小米、亚马逊的行动最为迅速、高效。任何企业，只有把握住了大数据时代的机遇、接受大数据时代的挑战，企业才能拥有了缩短发展时间、完成飞跃的筹码。

大数据和企业的生存发展息息相关，企业的管理层需要全面的数据源来确定正确的航向，全面的数据源搜索和分析需要专业人才，专业的人才需要经历商业气息的洗礼。这一切都成了企业在大数据时代直接面临的机遇和巨大挑战。得数据者得天下不再是一句标语，企业的整个商务链条都需要数据的支撑来保驾护航，失之毫厘谬以千里的教训时时刻刻都可能发生。重视大数据、对大数据了解详尽的企业高管才有可能带着企业在正确的路上，获得竞争优势。

进入大数据时代，让数据开口说话将成为司空见惯的事情，本书将从大数据时代的前因后果讲起，全面分析大数据时代的特征、企业实践的案例、大数据的发展方向、未来的机遇和挑战等内容，展现一个客观立体、自由开放的大数据时代。

目录

数据化带来的挑战前所未有

第一章 身处数据时代，揭开大数据的面纱

科技的迅速发展，互联网金融的兴起和繁荣，把数据推到了所有金融元素的核心位置。越来越多的企业逐渐认识到只有掌握正确的数据并看透数据背后的故事，才能够获得源源不断的财富。大数据时代伴着铿锵有力的节奏引领了世界的新潮流。

大数据到底是什么？

如果要追溯“大数据”这个专业术语最初的出处的话，就必然要提及apache org的开源项目Nutch。在那个时候，大数据的意思是更新网络搜索索引，同时还需要批量处理和分析大量的数据集。谷歌的Map Reduce和Google File System（GFS）发布了之后，大数据的定义中除了涵盖大量数据之外，还包括数据处理的速度。

研究机构Gartner曾给大数据（Big data）下过这样的定义：大数据是一种基于新的处理模式而产生的具有强大的决策力、洞察力以及流程优化能力的多样性的、海量的且增长率高的信息资产。

大数据一词源于英文的“Big Data”一词，以往也有类似的词语，如“信息爆炸”“海量数据”等等似乎都很难去准确描述这个词的具体内涵。麦肯锡全球研究所所做的《大数据：创新、竞争和生产力的下一个前沿》（James，2011）是这么定义“大数据”的：

大数据通常指的是大小规格超越传统数据库软件工具抓取、存储、管理和分析能力的数据群。这个定义也有很强的主观色彩，因为究竟什么样规格的数据才是大数据，这没有统一的标准，也就是无法确定超过多少TB（1000GB）的数据才是大数据。随着时间的推移和技术的发展，我们必须知道“大数据”的量会越来越大。还有一点，这定义也会因为部门的差异而发生标准的变化，这和通用的是什么软件以及特定行业数据集的大小有密切的关系。所以，现有各行业的大数据可以是几十TB，也可以是几千TB。

按照EMC的界定，特指的大数据一定是指大型数据集，规模大概在10TB。通过多用户将多个数据集集合在一起，能构成PB的数据量。

在IBM2011IOD大会上，负责IBM软件和硬件两大集团的高级副总裁Steve Mills曾说过："分析已经成为必要的能力，不再只是一个工具，是一种能让业务流程运转的智慧能力。企业要转化信息的洞察力为行动，而不是仅仅去争取竞争的优势，要将其转换为生存的根本。"

IBM公司概括大数据时有三个V，也就是大量化（Volume），多样化（Variety）和快速化（Velocity），此外它们还针对客户有了"大数据解决方案"的服务。IBM公司对大数据所概括出的三个V，其实也说明大数据潜藏的另一个V——价值（Value）。就这么说的话，大数据确实具备这四个V的基本特征。

大数据的第一个特征是数据的量大。电脑的数据运算和储存单位都是字节（byte），1KB（kilobyte）等于1024B，就是千字节。除此之外还有更高的单位MB（Megabyte兆字节），GB（Gigabyte，吉字节），TB（Trillion byte，太字节）、PB（Pet byte，拍字节），EB（Exabyte，艾字节），ZB（Zetta byte，泽它字节）和YB（Yotta byte，尧字节）。每一级之间的换算关系是1024。到了2009年，几乎每一个美国企业，只要是雇员人数超过1000人的，它的数据存储量大概都超过了200TB，这是十年前沃尔玛公司数据仓库存储量的2倍还多。在不少经济部门当中，企业平均的数据存储量甚至都达到了1PB。2010年欧洲组织的存储总量大概为11EB，这个数字几乎是整个美国数据总量

（16EB）的70%。2010年全球企业在硬盘上的数据存储量已经超过了7EB，而在PC和笔记本电脑等设备上的个人存储量也超过了6EB。美国国会图书馆当时存储的数据大概只是1EB的4000分之一（James，2011）。硬件技术的发展速度远远赶不上数据容量的增长速度，为此数据存储和处理的危机应运而生。巨大数量的数据被处理掉，例如医疗卫生提供商会将它们90%的数据给处理掉（这其中包括几乎所有在手术过程中产生的实时视频和图像资料）。

只不过，大数据不单纯只是大。海量数据存储危机的产生不仅仅是由于数据量爆炸性的增长，还有数据类型的改变带来的，这就是第二个V，多样化。此前的数据库用二维表结构存储方式就可以储存数据，譬如常见的Excel软件中处理的数据，这称为结构化数据。可是现在随着互联网多媒体应用的出现，像是声音、图片和视频等等非结构化的数据所占的比重在日益增多。有统计表明，全世界非结构化数据的增加率是63%，相对而言结构化数据增长率只有32%。2012年，非结构化数据在整个互联网数据中的占比已经超过了75%。

Informatica中国区的首席产品顾问但彬就提到过，大数据里有海量数据的含义，但它又大于海量数据的定义。简单来说，海量数据加上其他复杂类型的数据就是大数据的概念了。但彬还提到，所有交易和交互数据集都属于大数据，它的规模和复杂程度早已在依据合理成本和时限进行捕捉、管理和处理数据集的传统技术的能力之上。

简而言之，三种主要技术趋势汇聚成了大数据：其一是海量交易数

据，包括半结构化和非结构化信息，在从ERP应用程序到基于数据仓库应用程序的在线交易处理（OLTP）和分析系统的过程当中总在不断增长。企业很多的数据和业务流程也在不断走向公共和个人云转移，将造成更为复杂的局面。其二是海量交互数据。因为Facebook、Twitter、LinkedIn以及其他更多的社交媒体的兴起，这一部分数据诞生了海量的交互数据，其中涵盖了呼叫详细记录（CDR）、设备和传感器信息、GPS和地理定位映射数据，还有利用管理文件传输（Manage File Transfer）协议传送的海量图像文件、Web文本和点击流数据、科学信息、电子邮件，等等。其三就是海量数据处理。随着大数据的涌现，已经有很多用于密集型数据处理的架构应运而生，比如Apache Hadoop，它具有开放源码以及在商品硬件群中运行的特性。此外还有能以可靠、高效、可伸缩的方式分布式处理大数据的软件框架Hadoop。它之所以可靠，是因为它能够提前假定计算元素和存储失败，所以它能够维护多个工作数据副本，用并行处理的方式来加快处理能力和速度。Hadoop也是可伸缩的，PB级的数据它也可以处理。另外，Hadoop因为依赖于社区服务器，所以它的成本很低，不论是谁都可以使用。对企业来说，最难的在于如何通过成本效益的方式从Hadoop中存取数据。Hadoop最知名的用户是脸谱。通过Hadoop，像脸谱这一类的网站，也就可以自由地处理海量的数据，同时获得较高的收益。

“大”是重点，还是“数据”是重点？

先来做一个小测验。当阅读开始前，先停下来思考这么一个问题：

哪部分是术语“大数据”中最为重要的？是大，还是数据？还是二者都一样重要，或是都一样不重要？花一分钟的时间去思考这个问题。假如已经有了自己的答案，那就开始阅读接下来的内容。

既然答案已经有了，那就来看看哪个是正确的？显然，正确的答案应该是最后一个，事实上在大数据中，“大”和“数据”都不重要。其中最重要的是企业该如何去驾驭这些大数据。对大数据进行分析，以及在此基础上采取的业务改进才是最为关键的。

事实上，大数据本身是没有任何价值可言的。即便是一个人比另一个人拥有更多的数据，这也不代表什么。任何一个数据集，它们或大或小，本身都没有价值可言。如果不懂得如何去使用收集来的数据，那这些数据不会比地下室里的垃圾更有用。要是不投入环境或者是付诸使用的话，数据的意义就不在了。任何大量或是少量的大数据该如何体现自己的威力呢？要怎么去分析这些数据呢，又该如何去洞察或是采取什么样的行动呢？这些数据又要如何来改进业务呢？

很多人在阅读了众多炒作大数据的文章之后就相信之所以大数据要比其他数据有优势，就在于它的容量大、速度快和多样性，这种说法并不准确。在很多大数据当中，相比以往数据会存在更多毫无价值或是价值很小的数据。一旦大数据被精简到实际需要的容量时，它们所呈现出来的就不是大数据了。事实上这也不重要，无论是它被精简还是保持原本庞大的模样，这些关系都不大，最重要的是处理它的方式。所以说使用数据要比起它的容量更为重要。

大数据庞大的规模并非人们所关注的，包括它们能带来巨大的内在价值也非关注的事实。最大的价值还在于分析的方式，以及采用什么样的方式来改进自己的业务。

在人们阅读一本书的时候，关键点的第一个是大数据的大数据量，并且要承认大数据也是数据中的一种。只不过这并非企业兴奋的理由所在。这些数据使用时的新颖且强大的分析方式才是企业注意力集中的地方。作为社交网络应用的Facebook和微博，都构建了关联普遍用户的行为数据。人们在网络上浏览网页、购买商品、游戏休闲原本是不关联的。当智能手机推广普及之后，网络的行为越来越碎片化了。假设没有一定的关联，就很难去分析和利用这些数据。社交网络提供给用户统一的借口，让无论是玩游戏还是买商品的客户可以轻松地把碎片化的信息发到网络上。就像是一个用户行为数据连接器的角色一样，微博把所谓网络上用户的行为，完整地关联起来，画出一幅生动的网络生活图景，把用户的偏好、性格、态度等特征真实地反映出来，而这当中就是最为充分的商业机会。

彼此关联的数据价值要远大于孤立的数据。可是在当下数据孤岛是很常见的。个人计算机中的文件一般都会以某种类目来存放，内容和内容之间没太强的联系。企业之间也是如此，很多部门之间都壁垒林立，似乎每个人都愿意去保护自己的数据，从而形成“数据割据”的局面。只要是处在数据孤岛中，大数据所潜在的价值是很难被挖掘出来的。

与众不同的大数据

有别于传统数据源的大数据有不少重要的特征，不是每个大数据源都有这些特征存在，绝大多数的大数据或多或少地都存在一些这样的特征。

第一个特征是大数据的来源往往是机器自动的结果。人工不会干涉到新数据的产生过程，完全是机器自动的结果。如果拿传统数据源进行分析的话，就会发现它们的形成过程中会有人工的痕迹，像是零售业和银行交易、电话呼叫记录、产品发票等等，和某个人做的事情都有关系，无论什么情形，都会有人参与到新数据的形成过程中。可是大数据不是这样产生的，它不会在产生过程中与人互动，像是引擎中内置的传感器，即便没有人干预周围数据也会自动生成。

第二个特征是大数据作为一个全新的数据源，不仅仅是已有数据的收集扩展，比如在互联网中，顾客与银行、零售商之间可以直接在线交易。事实上这种交易方式和传统交易差异不大，不过是换一种渠道而已。企业通过收集网络交易数据就会发现这样情形下的数据和多年来他们得到的传统数据差异不大，不过是数量增加了而已。如果收集的是客户浏览行为的数据，那就会产生本质上全然不同的数据。

上面提到的相同类型数据，不过是数量多了的说法也会因为达到另一个极端，成为最新的数据，比如说传统读电表都是人工方式，也就是说自动读取用电数据的智能电表所产生的数据就是类型相同，不过是数量增加了。不过这种数据在某种程度上也能成为一种有别于人工读取的数据，应用更为深层次的分析方式，这样一来它们就可以称作是新的数

据源。

第三个特征是大数据中的大多数设计并非友好。实际上这些数据并未经过设计。就拿社交媒体网站上的文本流举例，用户不一定会被要求用标准的语序、语法和词汇表。人们的信息一经发布，社交平台就能够获得数据。这些不太规范的数据处理起来还是有一定困难的。在设计之初，大多数的传统数据都尽量要友好一些，就比如收集交易信息的系统最早生成数据会以整洁或是预先规范的方式来操作，这样形成的数据就更有利于加载和使用。还有一部分原因是由于要对空间进行高效利用，以避免出现空间不够的局面。

大数据有时候还会是凌乱和丑陋的。通常最开始传统数据就已经被严格地定义。每一比特的数据都存在重要的价值，这是必需的。一般大数据源一开始不会被严格定义，这和存储空间的开销越来越微乎其微有关，必须对各种有用的信息进行收集。所以说大数据分析的时候，各种凌乱丑陋的数据都有可能遇见。

最后的特征是海量数据并非有大量价值。实际的数据很多都是毫无价值的。在一篇网页日志当中，非常重要的数据就包含其中，当然也有好多没价值的数据也在其中。很有必要从中提炼最有价值的部分。定义传统数据源的起初就要求数据是百分百有用。这是因为可扩展性受到了限制，所以如果有没价值的信息在当中的话代价会非常昂贵。除了最初定义的有数据记录的格式外，数据内容和价值也被定义和约束了。当下存储空间的问题已经不存在了。大数据所收集的是所有的信息，然后再

去解决这些冗余信息所带来的问题。只有这样才会不遗漏所有的信息，与此同时在分析数据时的麻烦也会让人头疼不已。

大数据方式下的云计算

消费者会觉得大数据和云计算很无聊，可是对于Delphix来说却是一座宝藏，因为它正在利用这种技术进行敏捷数据管理。

Delphix不需要部署冗余的基础设施在自己的敏捷数据管理解决方案之上，还能同时提升流程的速度。客户因此能更为快捷地完成交付使用。其实敏捷数据管理就是企业数据库内虚拟化数据，再提高数据库驱动型应用的开发敏捷性质，因此使数据库和应用管理都发生大的改变。企业的数据库被Delphix放到了云上，再通过数据同步和虚拟化技术交给适当的人最恰当的数据。Delphix宣称有了应用交付解决方案后，应用项目的进度会提升5倍之多，成本会减少90%，事实上2010年Delphix面世后的销售增长率达到了300%。

成立于2010年的Delphix，2012年6月它的C轮融资就完成了2500万美元。这一次融资的领投是Jafco Ventures，投资人中还有Greylock Partners。迄今为止Delphix总融资金额高达4550万美元。公司依赖其“敏捷数据”拿到了超额认购。企业数据库的数据在“敏捷数据”的虚拟化作用下，增强了数据驱动应用的敏捷性，经济数据库和应用管理速度也提升了。

不少企业都把自己的目标设定为借由一个强大的平台来实现品牌推广，可是很多社交网站的数据还是找不到可行的商业模式，因为预期真

正得以实现的不多。不过社交数据公司在不断发展壮大，可以想见不远的将来社交网站的影响力利用问题不会再是遥远的梦想。

像是纽约的SumAll公司期望就是要带给每个客户“小而美”的数据。SumAll所提供的平台在于提供给中小企业实时的数据服务，利用桌面、iPhone和安卓系统来访问，可以看到很多可视性的大量数据，也就更便于阅读和观看。SumAll在和Shopify、PayPal和Magento合作电子商务和支付系统的时候，用户点击几下就能完成账户的集成工作。SumAll对于实时数据的分析很快速，再为用户提供一个如社交媒体式的“新闻订阅”一样的简洁分析和见解。SumAll还会为客户提供深入挖掘税收、发货和出售量的服务，甚至连对客户依照不同标准的排序分析也可以完成。

2011年11月成立的SumAll，在2012年6月著名风险投资公司Battery Ventures牵头联合Wellington Partners、Matrix Partners和General Catalyst Partners为SumAll投资了150万美元的种子期融资。SumAll到2012年12月对外宣布获得了600万美元的A轮融资，还是Battery Ventures联合Wellington Patners对其进行投资。目前设在纽约总部的公司有25名员工。

还有Ngdata公司，企业用户和他们的消费者通过它们能够进行一对一的营销模式提供和得到最好的建议和产品。Ngdata曾推出过一个产品Lily集成了内外部的结构化和非结构化的数据。Lily还可以用人工智能拍照工具对消费者的习惯和爱好进行记录。正在快速成长的大数据市

场，对企业的价值越来越大了，企业对市场的评估和行为的预判都要通过这些数据分析。ING的投资总监Tom Bousmans说过，消费者所产生的数据有上亿个，企业都可以通过这些来了解用户需求，彼此间还有个性和动态的互动。

成立于2009年的Ngdata的员工现有20名，它们还有类似Wibidata和Spire这样的竞争对手。Ngdata与竞争对手的不同在于它能够提供企业与消费者实现互动的数据解决方案，不仅是单纯专注在大批量数据分析之上。2012年10月Ngdata获得了250万美元的融资。这一次融资的资金主要来自ING、Sniper investment、Plug and Play Ventures等投资机构和一些天使投资人，这份资金将帮助Ngdata推广个性化产品线的拓展，并在纽约和旧金山专门为美国客户设立服务办公室。

Attivio的创始人Ali Riaz觉得企业用户每发送一条查询请求的时候，得到的信息都是具有洞察性的，绝非罗列出来的链接或是一张简单的图表。它回答的问题不仅是“是什么”还有“为什么”，就比如销售量下降是市场需求下降还是销售人员表现不够突出造成的。

任何一家企业要做的工作都是市场营销。近几年社会化媒体的兴起，让营销业者的注意力都集中在了数字营销之上，不过对于这个领域营销人员还欠缺有效的分析。Good Data公司正是瞅准这一商机，开始为营销人员提供集成服务，让他们可以利用微博等社交网络平台进行深度的分析。

大数据的奥秘

事实上并不是说大数据的处理就有多困难。收集一些数据，企业的分析专家团队就可以开始进行数据价值的探索。企业要做的就是要让分析专家团队最近地去接触那些数据，接下来的工作才是开始进行分析探索。要相信分析专家和数据科学家们都会很好地做好他们应该完成的工作。

一个很老的拇指法则指出，分析数据的工作的时间中有70%~80%都花在了收集和整理之上，剩下的20%~30%的时间才是真正的分析。大数据在刚开始处理的时候，分析时间所占的比例就更低了。一开始分析专家会用95%左右的时间来处理数据，甚至有时还会达到100%，接下来才会思考这些数据如何用来做更为深层次的分析。

上述做法必须得到理解。首先分析数据的流程中最重要的一环就是要先弄明白数据源的本质。对它们的表现、调整加载情况过程反复地进行检查，才能保证选择出能够更好地服务于目标的数据，虽然有很多数据并不那么吸引人或是令人兴奋，可确实是选出了最重要的数据。要是这些步骤都被忽略的话，那后面的分析环节就不能继续。

大数据中最优价值的部分被确定了，也就能确定用什么办法选择提取这些最优且精确的部分，这是最为关键的环节。这一环节必然是非常耗时的，尽管是花了比预计还要多的时间，也不用为此懊恼。因为企业的分析专家和业务赞助商是需要在弄明白数据源的过程中寻找一个代价最小、见效最快的方式。不论这东西是不是微不足道，但要给企业展示的一定是最有价值的东西。人们可以因此对此过程保持最高的兴趣，并

协助其他人理解进展的过程。

大数据要如何应用到业务中，这是个非常耗时耗力的过程。在这一过程中，企业的分析专家和业务赞助商总是想要找到最有效的方式，为企业展示这方面的进展，并以此来赢得最稳定的回报。

举个欧洲零售商的例子，公司总想获得最为详尽的网络日志数据。公司在经历了一个漫长且复杂的收集数据的过程后，开始实施了一系列简单的举措，其中包括鉴别用户浏览过的商品，在这些数据的基础上，他们建立了一个电子邮件系统，目的是为了给浏览过商品却未购买的顾客发邮件。企业从中获取了高额的利润。

公司采取了类似的基本早期措施外，还会继续投资于收集和加载网络数据之上。最为关键的一点在于他们缺乏处理整套数据流的经验和意愿。试想一下经过数据的深层次分析后他们所得到的回报。也正是因为这些快速及时的进展，每个在企业中的人都乐意再继续下去。他们或许从最早期的举措当中看到了数据的巨大威力，也已经为未来的努力买过单了。

顾客给予的是个体具体行为，这一点是最重要的。上述的例子当中已经让他们认定个人的重要性体现在了模式分析的输入数据上。每个个体从获取价值上来说并不需要识别。分析专家如今已经可以利用数据库技术在不识别个体的情况下完成分析。很多隐私的顾虑就可以就此消除了。事实上不少企业都这样来定位和分析具体个体的顾客。这些企业想必都已经制定了关于隐私保护的政策，不论在什么前提之下，这些隐私

政策都要谨慎地遵守。

世界在小数据时代是如何运作的，人们总是希望通过分析和收集数据来证明这一设想。不久以后，大数据时代的到来，人们就不再局限在这设想中。因为数据的分析和研究让人们发现了从前没发现的联系。

自然理论或是社会科学带来的人们的设想，也让我们理解和预测周围的世界。人类在经历假想时代到数据时代的过渡中，很可能会发现理论不再需要了。

《连线》杂志的主编克里斯·安德森（Chris Anderson）在2008年就指出："科学研究方法因为数据爆炸都变得落伍了。"此后他还在《拍字节时代》（The Peta byte Age）的封面故事中提到了，从某种意义上说，大数据就说明了"理论的终结"。安德森曾提出，如今不再需要一系列因果关系来验证各种设想了，这种传统研究模式早已被无理论指导的纯粹相关关系研究所代替。

安德森为了证明自己的想法，阐述了之所以量子物理学称为纯理论学科的原因正是由于高耗费、复杂实验等等。他所提到的潜在观点就在于量子物理学的理论已经和实际剥离了。他提到了谷歌设计的搜索引擎和基因排序工作，他认为："海量数据时代，所有其他的学科已经为应用学科所取代。只要是有足够的数据问题就可以说明清楚。假设有一拍字节的数据，并了解数据之间的相互关系，那问题就可以得以解决。"

安德森的文章在专业领域引起了轩然大波，事实上安德森也认识到了自己发出了过于偏激的言论，可是某种程度上说他的话还是值得深究

的。当中最为核心的是，截至目前，理论总是用来在实践中理解和分析世界的，可是到了大数据时代理论已经不再必要，数据就够了。也就是说譬如世界的运作、人类的行为等等普遍规则也都不重要了。现在最重要的是数据分析，一切问题都能解决了。

事实上，大数据也是有理论基础的。譬如分析大数据所用的就是统计和数学理论，还可能涉及部分计算机科学理论。尽管和地心引力理论差异很大，但不管怎么说也是理论。大数据分析模式如果建立在这些理论之上，那必然可以实现大数据的预测能力。事实上，为人们提供新鲜深刻的洞见就是大数据的价值所在。

第一要关注的是如何收集数据，是不是数据收集的方便程度就是决定的关键因素呢，还是数据收集的成本呢？众多的理论影响了人们的决定，就如达纳·博伊德（Danah Boyd）和凯特·克劳福德（Kate Crawford）说的那样，一定程度上，结果由选择所决定。谷歌用检索词来预测的是流感而不是鞋码。分析数据的时候选择工具依赖的同样也是理论。研究结果的解读同样应用的也是理论。在大数据时代不代表理论消亡，相反却是在大数据的方方面面都渗透着理论。

安德森作为第一个发现这个问题的人，尽管没有得到合适的答案，也值得让人们欢呼。大数据不会宣称“理论已死”，当然它也从本质上改变了理解世界的方式。以往的习惯都被颠覆了，旧有的制度也面临着巨大的挑战。

当人们选择大数据的理念和方法的时候，就会感觉大数据所释放出

来的价值已经让这些不再是一种权衡，而是未来的必然改变。高科技行业当中，不少人认为要依靠新的工具来到达大数据的彼岸，这观点可以理解为由于他们是工具制造者的缘故。大数据趋势的深层原因就是海量数据的存在，包括越来越多以数据形式存在的事物。

当下是大数据发展的最佳时机

迄今为止，在大数据上不少企业都做得不多。幸运的是到了2012年，即便是尚未重视大数据的企业也犹未为晚，当然电子商务行业的企业除外。不过这种情况很快就会发生改变。直到今天，绝大多数的企业所错过的不过是第一个吃螃蟹的机会，所以对它们而言这问题并不大，迎头赶上的机会还是存在的。不过过几年以后，要是这家企业仍旧不关心大数据的话，那它就会被淘汰。所以说，现在是控制大数据最好的时间点。

不管是什么企业获取业务价值都可以通过新的数据源来完成，而那种竞争对手还未发现这方式的情况不多见。大数据给所有人都提供了巨大的商机，无论是谁都要能打败自己的对手才能赢在最前方。未来的几年，人们会陆陆续续看到大数据分析所带来的成功案例。不少案例都会显示部分企业会毫无先兆地就被行业所淘汰了。而这些例子现在就已经引起了众多人的注意了，所谓这些企业正是那些在行业当中反应迟钝，落后守旧的企业。如果在新兴的电子商务行业中，则情况就有很大不同了。

因此，现在是最佳的时机。

实际上，控制大数据绝非想象中的那样难。不少大企业都已经开始了自己的数据收集和分析工作了，还将其视为自己发展战略中的核心部分。像数据仓库、报表和分析都已经应用得非常广泛了。企业只要意识到数据存在的价值，那么只需要延伸和扩展现有的工作就可以做到控制大数据。怀疑论者的话，诸如大数据没有探索价值，它们还没有得到验证，风险太大等等都不能信。过去的几十年，很多同样的借口也在一步步成为数据分析的障碍。还不确定大数据价值的人，必须让他明白大数据的分析和控制不过是现有企业所做事情的延伸罢了，并非本质性的变化。因此，大数据既然在人们身边，就不要害怕它的到来。

第二章 大数据如此重要，引无数英雄竞折腰

在这个一秒钟就可能江山易主的大数据时代，获得最真实可靠的数据并进行准确的分析和预测是企业占得先机的重要条件。如何在数据上展开角力成为企业生存的核心问题，只有企业对大数据的“外貌”和“内涵”有详尽的了解和把握，企业才有可能在未来的发展中分得一杯羹。

多样的非结构性数据

在电影《黑客帝国》当中，主人公尼奥在服下了蓝色药丸之后，就发现所有在他身边的一切其实都是数字化的幻想而已，他的工作、伙伴、住的高楼，看到的天空大地，甚至于他的情绪都不例外。电影的创作自然可以天马行空，真实的物理世界尽管不是如此，但不可否认的是它也在朝着数字化的方向高速前进。

像是高楼大厦，在动工之前就会形成一个涵盖了设计、施工、维护等多方面的综合建筑信息模型，它所使用的就是三维建模技术。在消费者看来，人们绝对会因为建筑信息模型的美观大方而自掏腰包购买效果图；在地产商看来，建筑信息模型所透露出来的信息便是他们需要为整个过程投入多少；在设计师看来，整个模型清清楚楚地呈现了所有设计的综合，他们能够在当中调整管线走向和通风设计等等；在工人看来，模型就是他们的施工图；在消防部门看来，即便是尚未完工的建筑也可以通过模型来评估它的消防效果，并模拟人群疏散的动态情形。总之，这建筑的方方面面实际上都已经数字化了。

日常生活中人们所接触到的文件、照片、视频，还有海量的数据，都有大量的信息蕴含其中。此类数据的特点是共同的，尽管它们的大小、内容、格式和用途并不相通。拿最为常见的WORD文档举例就会发现，最为简单的文档可能就只有几行字而已，但是一旦插进了图片、音乐等多媒体内容就可以成为一个多媒体的文件，文章的感染力就会增强。这一类数据就是非结构性数据。

结构性数据与之相对应，在结构性数据中人们对于表格中的数据可以简单解释，因为结构都是相通的。每个人每个月所领到的工资条，工资条的结构就没有变化过，变化的只是里面的工资和个税、保险。个人的工资条排列在一起就形成了工资表。结构化数据的计算机处理技术已经成熟了，会计和审计可以很有效地利用Excel工具来进行加减乘除、汇总和统计等一类的任务。要是有大量运算存在的话，商业数据库就会使用上，它们的任务就是存储和处理这些结构性数据。

可是，日常生活中无论是企业数据还是日常数据，大部分都是非结构性的。有咨询机构调查显示非结构性的数据占到了整个企业数据量的80%，还有调查显示高达95%，这个数据暂时还没有权威、准确的统计。信息产业这么多年一直在努力的方向就是让非结构性的数据能和结构性数据一样获得便利、快捷的处理。可是他们总在走弯路，一开始人们希望用处理结构性数据的方式来处理非结构性数据。只是非结构性数据个体之间的差异太大，用统一的处理模式来硬套的话，结果显然是不会太好。因此人们有很长一段时间认为非结构性数据的处理难度很大。

幸运的是谷歌公司成了大数据处理技术的先驱，它为公众提供搜索服务的同时，把大量网页、文档等数据的快速访问难题也解决了。雅虎公司也有一个研发小组，在谷歌技术的基础上成功地开发了一整套处理大数据的程序框架，这就是大众所熟知的Hadoop。目前这个领域的技术发展很是快速。

以上这些公司的技术研发，让不少人在面对非结构性数据的处理问

题上重新找回了自信，因此高清图像、视频等处理技术都进入了快速发展的时期。

社交网络上人们情绪表达方式也日渐丰富，企业为人们开发了众多表达心情的标准化图示，用以表达人们的各种复杂的情绪。

大数据的价值发掘

我们来说说银行、地铁中那些敏感部门或是地点的视频监控，凡摄像头的运转均为24小时，它势必会产生大量的视频数据。通常情况下的视频数据是枯燥乏味的，人们不会关心。但是一旦拍到了图谋不轨的行为，那么对于公安人员来说这视频就非常有价值了。可是事先人们不会知道哪一个部分有用，因此所有的视频材料都要保存下来，即便是存了一年的数据哪怕只有一帧对破案有用也是有价值的视频。不过对于研究人类行为的社会学家来说，这些视频都是非常珍贵的第一手材料，因为从中能发现人类的行为模式特点。

人们如今要获得医疗数据并非难事，手腕上的一块和电子表颇为类似的仪器就可以随时随地测量脉搏、体温和血压等等数据，再不断地将其传回医疗中心。数据除了能帮助人们检测自己的健康情况外，医疗保险公司也很是青睐这技术。保险公司的精算师依照这些数据的特点来研发新的保险产品，对他们现有的产品组合也是非常有帮助的。

上述的种种事例说明了：1.数据的价值是无可限量的；2.当然这价值犹如沙滩中的黄金一般需要挖掘；3.组合数据的价值要比单一种类的数据价值高得多。

在研究各行各业的数据应用中，会发现即使手中有一座如此大的宝藏，但挖掘工作仍是非常困难的，原因正是由于自身的数据中所蕴含的重生之道还不为人所知。互联网公司是最早意识到数据价值的公司，因此它们总在研究和分析领域领先。不过大数据的专利不再是属于大公司，它需要的是看待世界、产业的观念和视角。大公司通过它来合纵连横，扩张跨界，小公司也可以细水长流。关键问题在于如何看待大数据。

究竟多快才是快呢？

显然是小于1秒，就在分秒之间的客户体验。

传统数据应用和大数据应用之间的重要区别就在于此。十几年间，无论是电信还是金融行业都在经历着一场核心应用系统从分散到总部统一的过程。集中大量数据之后，所产生的第一个问题就是各类报表形成的时间延长了。业界在很长时间内都在质疑能否从海量增加的数据中快速地提取信息。

在这个领域，谷歌公司的贡献是有开创性的。谷歌的搜索引擎就仿佛在向信息业界宣布，全世界我那工业的搜索可以在1秒内完成，并得到所要得到的结果。大数据应用领域谷歌成了一个标杆。要是有超过1秒钟的数据应用的话，用户就会有不良的体验。下面举个营销方面的例子。

人们在购买越是昂贵的东西时就越是犹豫，会反复去掂量自己的购买能力。购买价格便宜的东西就越容易呈现出冲动购买的特征。根据消

费者的购买特点，京东商城将其分为了四种类型，其中37%是冲动购买者。对于这类购物者来说，能够在冲动的一瞬间为其送上最为精准的商品信息，是商品销售中的关键因素。幸运的是，关于这一点，社交平台的出现，为调查人们的偏好和兴趣提供了一个极好的平台，也让大数据时代这种精准的营销成了可能。

股票市场的交易主要是高频交易，要比他人快0.02秒才能有惊人的收益。为了能比他人快20毫秒，有人特地建了一条横跨西海岸到东海岸的光纤，还有人索性就留在了纽交所所在的街区。由毫秒时间差所造成的商业机会，此后会因为大数据的普及而出现在众多行业当中。

很多以应急反应为主的新兴产业很注重时效性。他们如果了解到某工厂有了事故，就会在第一时间做出判断，评估影响范围，到达现场并展开处置。

互联网投资创业现在的热点领域是O2O。经过商家门口的消费者如果能即时收到商家的促销信息，无疑是最为美好的服务。此时的促销消息若是消费者正好需要的商品或是服务，人人都能从中获益。消费者节省了时间，商家商品得到销售，服务商也获得了佣金。如果所提供的促销信息非准确时间获得的，那就会演变成为最为恼人的垃圾信息。谁都不愿意在任何时间任何地方收到垃圾信息，而这两种信息的差别常常只是几秒钟的差异而已。

数据的活性越高就有越大的价值。曾经有一家公司提供了数据样本希望有人能帮他们来评估一下潜在的商业价值。数据量很大，更新频率

也很高。这样的数据并非不常见，很多支付公司所收集到的交费记录常常都是如此。

数据的活性实际上就是数据的更新频率，更新频率越高的数据就有越大的活性，反之亦然。通常来说，数据集中的活性越大，就有越丰富的信息在其中。因此在大数据领域要有所成就的话，就要想办法去提高数据的活性。

对于公司的投资价值的判定，人们常常会听到这样的观点，公司是否拥有成规模和有活性的数据。之所以多样化和快速等特征不被提及，就因为人们更容易记住这一点。

大数据的结构化、非结构化、半结构化及多结构化

阅读大数据有关文章时，很多类似数据如何被结构化、非结构化、半结构化、多结构化的概念讨论会出现。通常大数据都是非结构性的，而传统数据是结构性的。只不过二者之间的差异并非绝对清晰。

应该说，大多数传统数据都是结构性的。这说明传统数据具有明确的、预先规范好的细节的格式等特点。任何一个时候出现的新数据都是以这样的模式呈现的。对股票交易来说，交易信息的第一部分格式就是月份/日期/年份等时间信息，然后就是12位的账户数字，接下来是3-5位字母所表示的股票代码。信息的格式事先就已经定好了，再由规范好的格式和顺序给出，这样的话处理起来就比较简单。

人们在面对非结构性的数据时，常常是没有控制权的，能做的就是接收它们。像是文本数据、视频数据、音频数据等等都是这样。图像是

独立像素用特定的方式组合起来的，只不过组合的方式是千变万化的。完全非结构性的数据就是如此。事实上大多数的大数据应该算是半结构化的。

半结构性的数据还是具备可理解的逻辑流程和格式，只是这些格式并非对用户都表现出了友好的姿态。半结构性的数据从某种程度上也可以被称作是多结构性的数据。此类数据，大量无价值的数据包裹着有价值的数据。相同情况下，理解和分析半结构性的数据难度要大很多。因此要用一套复杂的规则来理解半结构性的数据，只有在读了每一条信息之后才能动态地决定处理方法。

半结构性数据中最典型的就是网络日志。人们看到网络日志时，会觉得很丑陋，事实上它们中的每一条信息都有特定的价值。

非结构性的大数据源并不常见，反倒是大量的半结构性数据和多结构性数据比较常见。它们拥有可理解的逻辑流程，从中能提取出能用于分析的信息。只不过要驾驭半结构化的数据，必须有时间的保障，这才是处理它们的最佳方式。

网络日志的信息是有逻辑的，尽管很多时候很难看出来。日志中字段和分隔符都是不同的，仿佛是结构性的数据一般，包含了大量的价值。只不过它们彼此间非以固定的方式相联系。不同网站上的网络日志点击所用的时间长短不一。另外，半结构性的数据要理解其内在的逻辑并非不可能的，只不过要花上一段时间才行。

分析专家似乎更怕非结构性数据，半结构性的数据对他们来说，要

征服是需要付出努力的，而他们也能够做得到。分析专家首先要把半结构性的数据组织成结构性的，然后再在他们的分析流程中运用。而对于非结构性的数据来说，困难就会大很多。即便是已经征服了半结构性的数据，他们要面对非结构性的数据时还是会感觉是个巨大的挑战。

大数据是扩展性的下一代传统数据

关于大数据的各种言论作为为众人所关注的热点层出不穷，核心还是大数据要如何从根本上变革分析和使用大数据的方法。仔细思考一下，会发现事实并非如此。

关于庞大的大数据和它们的可扩展性已经不是个新鲜的话题。第一次使用新数据源的时候，通常会认定大数据是庞大的且很难使用的。事实上大数据不过是突破了当前极限，更大规模的数据罢了。分析专家对于传统数据源的掌握即便是有瓶颈的，但这也不影响其驾驭大数据源。毕竟分析专家从一开始就非常努力地在探索新数据源，并且将持续下去。

在电信公司中分析电话详单的第一人是谁？当然是分析专家。深入研究零售点销售数据并为此找到价值的第一人又是谁？当然也是分析专家。最初，分析几千个到几万个，甚至几十万个商品的数据被视为是一个巨大的难题。可是现在这一切已经迎刃而解。

最早开始做数据源分析的分析专家处理的数据便是在当时被认定为无法处理的数据。首先他们要做的是找到分析和利用的方法，并且是在当时的条件下。不少人对可行性产生质疑，还有人认为这些数据并非具

有价值。

分析专家们正努力在做的事情和他们做事情的理由不会因为大数据而发生改变，即便从一开始很多分析专家会自称为数据科学家，可是他们的目标还是一致的。待解决的很多问题都涉及了大数据，这跟从前差别不大。最后分析专家和数据科学家们还是会去探索一直以来探索的事情的，并从中发现一些有价值的趋势和模式。必须强调的一点是，尽管大数据听起来很陌生，可是它所带来的挑战却无须畏惧。

不论从什么方面来说，大数据所产生的问题一定是从前出现过的。在数据分析领域，永恒的主题就是驾驭新的、可扩展性极限的大数据源。大数据不过是下一代的这种数据罢了。处理此类情况，分析师已经很是熟悉了。只要是有驾驭过其他数据经验的企业，大数据也同样可以驾驭。

分析专家的工作策略有一部分会因为大数据而改变。他们常常将新的工具、方法、技术和传统分析工具相结合，目的是为了有效地处理大数据流。要从中提取最有效的信息，复杂的过滤算法是十分必要的。当然也要更新建模和预测程序，大量的数据要输入添加到现有的输入当中去。

从根本上来说，分析的目标和流程并不因工作策略的改变而改变。大数据催生的是创新性的分析方法，分析专家也必须在继续扩展性的瓶颈下革新。可是必须承认的是，分析专家在处理大数据和以往数据时的方式差异不大。

是什么构成了大数据价值链？

依照所提供价值的来源（数据本身、技能和思维）不同，大数据公司可以分为三类。

第一类是在数据本身基础上建立的公司，它们拥有海量数据，或者说至少可以收集海量数据，但在提取数据价值上和催生创新思想方法并不是最佳的，例如Twitter。它拥有的数据是海量的，只是这些数据还都要通过授权给其他两个公司来供他人使用。

第二类是技能型的公司。一般来说它们是咨询公司、技术供应商或是分析公司。它们有专业的技能，可是却不一定拥有大量的数据以及提出创新性使用数据的才能，例如天睿公司（Teradata），它就是一家大数据分析公司，而它的营销电子所用的数据都是来自沃尔玛和Pop-Tarts这两个零售商。

第三种是思维的公司。这种公司中的代表就是Jetpac。Jetpac的联合创始人皮特·华登（Pete Warden），就是一个通过想法来获取价值的最佳例子。利用用户分享到网络上的旅行照片，Jetpac寻找人们下一站的旅行目的地。很多时候数据和技能并不是某些公司成功的关键，真正让他们脱颖而出的是其创始人和员工的创新思维，以及如何从数据中将价值中挖掘出来的独特想法。

谷歌的首席经济学家哈尔·范里安（Hal Varian）也认为，世界上最棒的职业就是统计学家，这个说法非常有名，他提到："要是想成为成功者，那必须是稀缺的，不可替代的人，绝不能是普通的，随意可以被

替代的。”他还说道：“数据多且有战略上的重要性，可是能从数据中提取价值的能力却不是人人都具备的。这也是为什么统计学家、数据库管理者和掌握机器理论的人是最了不起的人的理由。”

技术和技能的过于强化会让数据本身的重要性被弱化，因此这是不可取的。计算机行业的快速发展，落后人力技术会渐渐远去，而范里安所赞许的技能就会开始普及。

大数据中最有价值的部分就是它本身，所以智者会先考虑如何拥有大数据。尽管他们并非第一个获得这些数据的人，但他们能和接触到数据、有权使用数据或是有意将数据授权于他人的人有接触。

一部分企业就巧妙地把自己放在了信息链的核心，于是扩大规模和挖掘数据的价值就成为可能。信用卡行业就符合这一点。近些年来，很多中小银行因为防范信用诈骗的成本过高，总是不愿意发行自己的信用卡，发行信用卡的都是大型金融机构，它们才能负担起防范技术所需要的人力物力。美国的第一资本银行和美国银行就积极地承担了这一工作。只是到现在，中小银行没有自己发行的信用卡，它们对于客户的消费模式全然不知，也就无法为客户定制特定的服务，关于这一点它们已经后悔了。

Maste Card长期就处在这个信息链的中心，因此数据收集和数据价值挖掘的黄金位置为之所占据。可以想见，未来信用卡公司不会再收取交易的佣金，反倒是会提供免费的支付服务。因为从中它们可以获得大量的数据，在经过复杂的分析之后，它们能够依靠卖掉分析结果来获得利

润。

来说说第二类公司。同样是Maste Card，它们有自己的分析系统，所以它是游离在第一类公司和第二类公司之间。不少公司还会选择发展其专业技能，例如埃森哲咨询公司，它就通过和不同行业公司合作，应用高级无线感应技术来收集各类数据，并对此进行分析。

在医学数据领域，技术公司如何提供有效服务的例子比比皆是。华盛顿州的华盛顿中心医院这么多年来就和微软研究中心合作分析了众多的匿名医疗记录，这当中包含了患者人口统计资料、检查、诊断、治疗资料，等等。此研究的目的在于降低感染率和再入院率，要知道以往医疗卫生领域开销最大的就是这两项。无论是哪一个项目比例降低了都意味着开支的巨大降低。

研究中发现了不少惊人的相关关系，不少情况下出了院的病人一个月内又会再次入院。这当中就有不少是已经很常见却始终没有好的解决方案的，譬如，患有充血性心力衰竭的病人就有再入院的可能性，这病的治疗困难很大。研究当中还发现有一个因素是很出乎人意料的，那就是病人的心理状态。病人若是对最初诊断中有类似“压抑”这心理暗示疾病的词汇时，病人就有很高的再入院可能性。

尽管这种相关关系很难推导出特定的因果关系，但这也说明病人在出院以后要以解决心理问题为医学干预的重心，这显然对身体健康有着重要的帮助，也对降低再入院率，提供更好的健康服务，降低医疗成本更有利。机器在一大堆的数据中筛选出了这相关关系，人类或许永远也

发现不了这个问题。这些数据仍属于医院，微软只提供了分析工具，依靠Amalga系统来帮助发现有价值的信息，这并非什么出彩的想法，却是这里最需要的。

技术专家是挖掘数据价值的专用人才。他们在被赞扬的同时还获得了“数据武士”这样一个非常时髦的称号。事实上他们并非想象中的那般耀眼。他们在大数据当中淘金，最后却要将财富拱手送给大数据的拥有者。

第三种类型的公司是拥有大数据思维的公司和个人。先于他人发现机遇是他们的优势所在，当然他们本身未必具备专业技能或是数据，甚至有可能他们还是行外人，可正是因为这个原因他们的思维才更为活跃。他们要考虑的仅仅是实施的可能而非可行性。

克罗斯和他的Flight Caster是这一类公司当中最先有行动的一个，尽管也没比其他企业快多少。大数据思维本质上是一种意思，它认为数据公开后如果能得到恰当处理的话，千万人急需解决的问题就会得到解决。Flight Caster在2009年8月公开发布了。就在同一个月，Fly On Time.us的计算机专家们也纷纷投入到公开数据中去，并建立了自己的网站。很快，Flight Caster原本的优势被削弱了。到2011年，克罗斯无奈之下只好把公司卖给了Next Jump。

随后克罗斯的重心就转向了新闻行业，很多人将其视为夕阳行业。在他眼里，并非如此，新闻行业是一个创新型外行人可开垦的宝地。他创办了新的创新科技公司Prismatic开始收集网上资源并排序，所谓的排

序是在文本分析、用户喜好等基础上建立起来的。最重要的是，这个系统包容性很强，无论是青少年博客，还是企业网站，只要是受欢迎的内容就会排在非常前面。而欢迎与否的判定标准则来自点击率和分享次数。

Prismatic显然在服务上非常关注青年人和媒体交流的新方式，所谓信息的来源被弱化了。这无疑给那些自视清高的主流媒体提了一个醒，公众的力量已经超越了它们。最让人惊讶的是，从新闻领域内部诞生的Prismatic手握着大量的数据。美国国家记者俱乐部的那些“老人”对于网上的媒体资源始终没关注过，甚至连阿蒙克、纽约和印度班加罗尔的分析专家们也从来没想过这种方法。一头蓬松头发的克罗斯，说话吞吞吐吐，可就是这样的一个外行人，有想法也做到了，他的数据告诉了这个世界，相比《纽约时报》还有更为有效的信息来源。

同上世界90年代相比，大数据思维以及拥有此思维的人和初期电子商务的情况大相径庭。最初电子商务的从业者不受传统行业的固有思维和制度缺陷所限制，所以当时还在对冲基金工作的金融工程师杰夫·贝索斯辞职办了一家名为亚马逊的网上商店，软件开发工程师皮埃尔·奥米迪亚（Pierre Omidyar）开发了一家拍卖网站。现在若是领导者是具有大数据思维的，那他们一般不直接拥有数据资源。正因如此，他们不会在利益和金钱欲望面前让自己的想法实践受到阻碍。

正如人们见到的那样，也有一部分公司集合了大数据的大多特点。相比他人，埃齐奥尼和克罗斯决胜的想法不但早人一步，同时还存在技

术上的优势。Teradata和埃森哲的员工上下班很是守时，不过他们也常常有灵光一现的时候。这一切都是由于公司所承担不同角色而造成的。很多公司会将自己所掌握的数据以授权的方式授权给更有能力挖掘数据价值的人。

现在不少大数据的先驱者通常学科背景都比较复杂，当自己掌握了数据技术之后就会和自己的背景知识相结合，然后在广泛的领域中应用。天使投资人和企业家的新一代正在诞生，他们中的大多数来自谷歌离职的员工以及所谓的Paypal黑手党”。这些人和一部分计算机科学家成了当下众多数据科技公司的最大依靠。

大数据只要得到有效的利用，公司的盈利模式和传统交流方式就会得到改变。举一个非常典型的例子，欧洲一家汽车制造商就利用了获得对手所没有的行业信息，而重新定位了自己和零件供应商之间的关系。

现在的汽车通常都配备了芯片、传感器和各种软件，汽车启动之后，这些关于汽车的信息就会反馈到制造商的电脑上。一部中档轿车的微型处理器大致为60个，车辆总价值的三分之一是车上的电子仪器。汽车上装满了这一类的车载仪器也使得车子成了“漂浮的观景台”。汽车零件的工作状态通过这些电子设备来监控，整体整合之后可以提高汽车的运行质量。总的来说，一个能够掌握这些数据的公司必然在竞争中占得先机。

汽车制造商和行业外的数据分析公司合作，结果显示产自德国的油箱蒸汽泄漏检测传感器存在一定的故障，即使是好的油箱它也会发出多

达16次的报警报告。得到这一消息后，制造商反馈给供应商。商业环境比较和谐的情况下，就会出现已经在这个项目上投入大量资金的制造商，利用这一部分数据来为自己挽回一点点损失的情况。

制造商要考虑的是接下来该怎么办，是卖掉这些数据吗？它的价值有多少？假设供应商推卸自己的责任怎么办？或者是自己在挖掘数据中有了失误？制造商很明白，信息一旦被公布，其他的竞争对手也会有的放矢来改进自己的产品。所以比较明智的做法应当是只能让自己从中获益。最后汽车制造商找到了一个完美的方案，那就是利用软件的改进来改进零件，然后再为此申请一个专利，最后再将专利转卖给供应商。

大数据时代真的来了

继美国政府推出“大数据研究和发展”计划之后，日本政府又重新启动了ICT战略研究，此研究曾在大地震时期暂时停摆，这是一个重视大数据应用的战略计划。联合国此后也发布了《大数据促发展：挑战与机遇》白皮书，全世界似乎都在迎接大数据时代，各种计划接二连三地发布。

日本总务省信息通信政策审议会下设的ICT基本战略委员会在2012年5月召开会议会。会上，大数据研究主任、东京大学的教授森川博之提到，在大数据技术领域美国的优势是明显的，像是谷歌、亚马逊这样的大企业都在大数据的应用领域拥有很强的技术优势，日本接下来必须在大数据方面制定一系列战略来应对大数据时代。日本文部科学省在7月就发布了以学术云为主题的讨论会报告，提出大数据时代学术界要做好

迎接挑战的准备，主要在大数据收集、存储、分析、可视化等等阶段展开研究，并构建大数据利用模型。

联合国2012年发布的《大数据促发展：挑战与机遇》白皮书已明确提出大数据时代已然到来，对于联合国和各国政府来说，这是一个历史性的机遇。报告中还对政府如何利用大数据来响应社会需求，指导经济发展进行了讨论，提出要在联合国成员国建立“脉搏实验室”，主要用于挖掘大数据的潜在价值。澳大利亚出资赞助印度尼西亚政府在其首都雅加达建立了“脉搏实验室”，于2012年9月投入使用。

大数据当前还是个新兴前沿的概念，我国尚未从国家和政府层面提出大数据相关的战略，可是在2011年11月，工信部发布的了物联网“十二五”规划中明确提到了四项关键技术创新工程，包括了信息感知技术、信息传输技术、信息处理技术和信息安全技术，当中的信息处理技术就有海量数据存储、价值挖掘等等方面的智能分析技术，显然这都是和大数据密切相关的技术。也就在同时，广东省等地方政府已经率先启动了大数据战略，推动本省的大数据发展，协助开放共享。

作为国家的核心资产，各国已经开始了大数据的激烈竞争。一旦在大数据领域落后，必然就无法守住本国的数字主权，也就意味着难以占据产业战略的制高点，国家安全数字空间也会相应地出现漏洞。美国政府在大力推行“大数据研究和发展”的计划之下，欧盟、中国等大型的经济体也会在不久的将来出台属于自己的引导性和倾斜性政策，目的就在于抢占大数据的战略制高点。一轮关于大数据的新竞争马上就要登场。

历史上这样的一幕曾经出现过。1993年，美国出台了“信息高速公路”计划，各国因此反应十分强烈。同年日本政府发布拟建设“研究信息流通新干线”计划，将全国的大学、研究机构利用高速通信线路来连接，并在后一年的5月又提出了日本版的“信息高速公路”计划，前后发布了《通信基础结构计划》和《通向21世纪智能化创新社会的改革》两个报告，报告中对网络建设的实施分三个阶段进行。欧盟在1993年6月的哥本哈根欧盟首脑会议上，由当时的主席德洛尔首次提出了“构建欧洲信息社会”的倡议，之后又在12月发布了旨在“振兴经济、提高竞争能力和创造就业机会”的白皮书，白皮书中已经提出了欧洲版“信息高速公路”构建的清晰构想，还为此成立专门的工作小组主要负责推进整个计划。与此同时，加拿大、韩国、新加坡等发达国家也都在逐步开发自己的技术优势，只为占据高新技术的制高点，迎接21世纪到来的技术发展挑战。各国都不惜投入巨额资金推出各国版的“信息高速公路”计划，一时间全球范围内“信息高速公路”计划风生水起。

从本质上来说，大数据就是人类社会所有数据量变到质变的必然产物，是“信息高速公路”计划的进一步升级和扩展，它对人类社会未来的走向和发展势必会有巨大的变革意义。很显然，现在的趋势已经说明了大数据时代真的到来了。

第三章 需求挖掘与分析，电子商务与大数据

互联网购物逐渐成了人们购物的主要方式，电子商务的繁荣带动了相关产业的不断升级。为了迎合大数据时代的特征，也为了更好地把握住大数据时代的种种商机，各个电子商务企业纷纷涉足互联网金融，追逐大数据，都希望占据未来市场。

大数据时代中的电子商务

全球最大的信息技术和业务解决方案公司IBM在被联想收购之前，可谓是电子商务行业中的一匹宝马。它始创于1911年，总部位于美国纽约州阿蒙克市。IBM在2011年的净利润达到159亿美元之多。

进入大数据时代之后，IBM积极响应并做出转变。IBM全面整合了公司的内部资源，搭建了全新的数据平台。就此宣告全面升级的大数据战略。IBM的大数据战略体现在三个方面：1.包括掌控信息、获悉洞察、采取行动的全面战略理论（也称3A5步）；2.包括Hadoop系统、数据仓库和信息整合系统、流计算的全面的解决方案；3.全面的落地实践。

马云的阿里巴巴可谓是国内电子商务界的龙头老大，阿里巴巴成功的因素之一就在于该企业非常重视数据。1999年成立的阿里巴巴经过十几年的发展，企业平台上已经积累了大量的数据，目前阿里巴巴拥有阿里巴巴B2B、聚划算、一淘、淘宝网、天猫商城、中国雅虎、阿里云、一达通、中国万网等子公司。

阿里巴巴设立了“首席数据官”一职以充分挖掘大数据的价值，并为自己的网络销售平台提供完善的数据云服务。在2012年的商业大会上，马云表示阿里巴巴将于2013年1月起转型金融、重塑平台和数据三大业务。阿里巴巴希望通过这种方式，分享和挖掘海量的数据并为其他中小企业提供更有价值的信息。

除了IBM、阿里巴巴之外，甲骨文、微软、惠普、百度等公司为了

应对大数据时代的挑战和机遇，都在积极挖掘大数据之中的“宝藏”，使得大数据市场一时间热闹非凡。而在这热闹的过程中，众多的电子商务企业在大数据的基础上，开始纷纷涉足互联网金融。网络银行一夜间崛起，网上支付、手机客户端支付已经成为很多人消费的主要支付方式。京东商城选择和中国银行合作，担任着类似的中介角色。通过这样的方式为供货商们提供入库单融资、应收账款融资、订单融资、资产包转移计划等服务。京东商城对供应商提出的融资申请进行核准之后，转交给银行，银行根据相关材料对供货商发放资金。

2013年1月，京东商城的CEO刘强东在公司内部年会上表示，京东商城将以大数据的供应链金融业务为主要战略性业务，以此应对大数据时代的考验。此外，京东商城还将组建京东金融公司等子企业和部门为商家和个人提供融资贷款服务。

阿里巴巴和京东商城这两个老对手，在进入大数据时代之后交手更加频繁。京东并不是首家涉足供应链金融的商家，阿里巴巴在2007年就已经推出了供应链金融以帮助中小企业进行融资。供应链金融的好处和利润被其他各个企业看中，这些小企业争相挤进这个领域，想要分得一杯羹。

其中最具代表性的就是金银岛。2009年，金银岛通过和中国建设银行、中远物流合作推出了E单通。E单通可以细分为网络订单融资和网络仓单融资两部分。这是合作三方共同建立的一整以实现物流、资金流、信息流的深度融合为目的的服务和风控体系。

亚马逊在大数据时代的实践

亚马逊是全球第一家网络书店，他的掌门人是贝索斯。了解亚马逊的人都会发现，贝索斯有个习惯——在开会的时候留出一把空椅子。他的目的是为了告诉自己的员工，消费者是公司最重要的人。本着顾客至上的经营理念，亚马逊成了名副其实的电子商务领袖：仅2011年的净利润就达到了6.3亿美元，公司的市值接近千亿美元。

和很多企业不同的是，亚马逊推出新产品和服务的速度非常快，并不需要进行长时间的调研、分析等步骤。因为只要推出了新产品，几周的时间内，消费者就会自动为公司留下珍贵的评论及购买的数据。

有了这些数据，亚马逊就可以采用大数据技术，详细分析这些数据背后的真相。准确地评估自己的新产品，从而很快给出产品是否应该继续推广、如何推广等问题的答案。亚马逊的这个流程，实际体现了以消费者为中心的理念。通过消费者的真实反映来改善产品，这是很多企业都应该学习的。贝索斯的“空椅子”理念代表了他对电子商务的深刻理解，也代表了亚马逊在大数据时代的良好实践。

在金融市场不断发展和稳定，金融产品在人们的生活中不断占据更大的比例的情况下，全球已经进入了从工业化时代向信息化时代过渡的阶段。美国是这条道路上走得最快的国家，中国紧紧地跟在西方发达国家之后。

以宝洁、沃尔玛、通用汽车等企业为代表的大工业时代有明显的特征——大量生产、物流强度大、零售量大。大工业时代不会在未来的消

费市场延续，取而代之的是信息化时代。信息化时代的特征有：主导者就是消费者、个性化生产、网络化协作。

这一转变将一改企业只专注于内部管理、生产和供应链的问题，使得消费者处于主导的地位并成为企业的一分子。

以著名的汽车生产商福特为例。福特在早期一直生产T型车，并且福特的理念是以生产为中心，当时福特对外流传这样一句话“顾客可以随心选择自己喜欢的颜色，只要是黑色”。但是这样的经营理念，早已经被淘汰。在当下的市场竞争中，企业必须学会揣摩每一个消费者的喜好，以提供满足绝大部分消费者需求的产品。

开启了各个企业以消费者为主导模式之门的大数据时代到来了。比如一家服装企业企图调查其顾客的购买意愿，于是安装了监控，以方便观察顾客试穿了哪些衣服。但是既要安装设备，又要整理和分析数据，成本一下子提高了，尝试以减少样本量的方式降低成本只会使得统计的结果有偏差或者失灵。

如果将类似的情况放在互联网上，边际成本的增加却趋近于零。因为消费者只要在网上点击了衣服的图片、放进了购物车、付了款等等，这些动作都会被服务器自动记录下来。想要分析这些数据唯一花费的就是时间，从海量的数据中挑选需要的数据并做出分析。这一对比，大数据技术的价值就明显地展现了出来。

想要成功地实现展业转型，就必须洞悉消费者的心理、快速响应消费者的需求，即使是潜在的需求。企业如果拥有了对大量数据的处理和

分析能力，就等于拥有了“千里眼”和“顺风耳”。虽然大数据技术在国内很多企业中都处于起步阶段，但是总有一些优秀的领袖企业，带领着大家走在技术进步的道路上。

以对比亚马逊和索尼为例来体现亚马逊对待大数据技术的态度和实践。亚马逊拥有全球最大的电子商务网络，这也使得亚马逊可以大规模地收集消费者在浏览和消费的过程中留下的一切痕迹，从而拥有全面精准的数据。以此为中心，亚马逊开发了市面上唯一一款能够和苹果的iPad抗衡的产品——kindle fire。

而索尼的市场份额在近几年不断缩小，这是因为它所生产的产品明显和消费者的喜好脱节，walkman等产品早已经被消费者淘汰了，而索尼似乎后知后觉。这正是因为索尼错过了利用大数据技术更紧密地贴近消费者的机会。

美国明尼阿波利斯市的一个经典案例一直在互联网上流传。塔吉特百货里来了一个怒气冲天的男子强烈要求见百货公司的经理。原因是百货公司寄给了她在上高中的女儿一些购买母婴用品的优惠券。顾客因为百货公司不恰当的行为非常生气。

经理查看了公司发给顾客女儿的邮件，的确给他的女儿赠送过购买母婴产品的优惠券。经理反复和这位顾客道歉，才平息了这件事情。

有趣的是，几天之后百货公司的经理再次给这位父亲打去电话表示歉意的时候，这位父亲非常愧疚地告诉他：“对不起，我之前错怪了你们，我和女儿长谈了一次，她的确怀孕了。”

很多人看到这里都会有疑问：百货公司怎么会知道顾客的女儿怀孕了，并向她赠送母婴产品的优惠券呢？商店竟然比顾客先一步了解他女儿的动态，这实在令人惊讶。想要知道这其中的奥妙就必须提起电子商务的鼻祖——亚马逊。

成立于1995年的亚马逊是全球最大的电子商务供应商。该公司最令人称奇之处在于，在成立之初，亚马逊一直处于亏损的状态，而且逐年严重。数据显示，2000年亚马逊亏损了14.1亿美元。在成立之初的8年时间里，亚马逊一直处于这样的亏损状态。直到2003年，亚马逊才终于开始盈利。

亚马逊的“锦囊”就是它有一位伟大的领导者——贝索斯，贝索斯是一个眼光长远的人。尽管在成立初期，亚马逊经历了互联网泡沫的冲击、经历了投资机构的做空，贝索斯依旧不顾外界对公司的评价，我行我素。在贝索斯给公司股东的信件中，他总是强调：“It’s all about long term.”而亚马逊之后的辉煌也证明了这位领导者的犀利眼光。

仅仅2012的前三个季度，亚马逊的营业收入就达到了398亿美元之多，相比前期，涨幅非常大。亚马逊能够做到如此转变主要归功于它强大的推荐系统。亚马逊的网店系统最强大的一点，就是能够让顾客发现自己的潜在需求。顾客进入亚马逊的网上商店起，就不断地被这样的思想渗透。“人气组合”“购买了此商品的用户还浏览了”等等栏目都是吸引顾客发现自己潜在需求的“钓钩”。

亚马逊将顾客在网站内的所有行为都通过系统记录下来，根据数据

的特点进行分类处理，按照商品类别形成不同的推荐栏目。例如“今日推荐”就是根据当天顾客浏览的信息记录，推出一些点击率最高或者购买率最高的产品。而“新产品推荐”则是根据顾客搜索的内容为顾客提供了大量新产品的信息。“用户浏览商品推荐”，则是将顾客曾经浏览过的商品信息再一次推向顾客，让顾客考虑购买或者进行二次购买。

捆绑销售法也是亚马逊采用的有效方法之一。即利用数据挖掘技术分析顾客的购买行为，找到某件商品的购买者经常一起购买的其他商品，构成销售组合，进行捆绑销售。他人购买或浏览过的商品栏目，则是通过社会化的机制，根据购买同类商品的顾客的喜好，为顾客提供更多的选择，使顾客更加方便地挑选。

亚马逊成功的另一个主要因素是，在给顾客做推荐的时候，亚马逊的顾客体验非常特别。亚马逊给出了大量的真实数据，让顾客在短时间内就对其产生了信任和忠诚。例如，购买了此产品的顾客还购买了某产品、你曾购买过某产品因此给你推荐类似的其他产品等等。

不得不提的是，亚马逊的推荐内容是根据真实数据分析计算出来的。每个用户的档案中都记录了该用户的所有购买和浏览行为。亚马逊的商品评价系统也从另一个方面归纳和反映了顾客对产品的偏好。亚马逊将这些数据做成顾客的档案，直截了当地告诉顾客亚马逊这些推荐举动的可靠和用心。

小米手机在大数据时代的实践

小米手机就像是竞赛中的一匹低调的黑马，在赛前默默无闻、鲜为

人知。一下子就占据了手机市场上前列的位置。于2010年成立的小米，自2011年推出第一代小米手机起，就获得了顾客的喜爱。2012年小米又推出了第二代手机，并且开通了网络销售平台。令很多竞争者没有想到的是，网络销售平台一开放，第一轮的5万部手机在3分钟之内被抢购一空。到2012年中旬，小米公司的市值已经达到了40亿美元。小米的董事长雷军推崇软硬一体化设计，和苹果公司的创始人乔布斯相同，因此，很多小米的喜爱者都称雷军为“雷布斯”。

虽然“雷布斯”的小米公司在很大程度上是模仿乔布斯的苹果公司。但是一味地模仿是不会取得这么大的成功的。小米有着自己的一套“模仿”模式，这种模式分为三个方面，即硬件、软件、移动互联服务。小米的目标是要把这三个方面都发展到市场领先地位。

这个年轻的公司，在其走向市场之后，已经发布了小米1、2、3三代手机。其软件上的实力也不可小视，小米公司开发了米柚操作系统、小米读书、小米分享等软件。

这样一家看似模仿苹果公司而产生的企业，在其成长和发展道路上一定受到了很多的质疑和责难。但是小米却用不断提供好的服务来回馈了这些。小米选择的道路是正确的，同时掌握硬件和软件的技术，为顾客提供全方位的服务，其未来一定会受到越来越多人的认可和喜爱，发展前景不可限量。

再好的软件，没有硬件作为展示的平台，也无法发展起来。就好像一个人没有了灵魂，生命也索然无味一样。只有将硬件和软件结合起

来，提供全方位的服务，才能给顾客最好的体验。科技迅速发展，电子产品不断升级的时代，顾客不仅重视电子产品的外观，更重视其内在的软件配置和操作系统。一些专攻硬件的制造商，如果不转变自己的生产方向，那么就会在竞争中被淘汰。

任何一家电子产品的生产商，如果没有自己独立的软件配置，都无法在竞争中生存下来。很多生产商在自己的产品中内设自己独有的软件，来保护自己的产权。例如，联想的手机就默认联想的网盘，华为的手机默认华为的网盘。可见，硬件和软件同样重要，缺少一个，就会被竞争对手获得先机。苹果公司的成功就极大地证明了软、硬件一体化的优势。

软硬件一体化的模式，适用于各种智能型电子产品的开发，而不仅仅是手机生产。不论是笔记本电脑还是iPad等平板电脑，都遵循这样的模式。小米的发展空间远不止如此，它可以发展成为一个综合性多元化的电子产品生产商。通过将这种模式扩展到小米电视、小米电脑等产品的开发和生产上，不久的未来，小米还将带给我们惊喜。

没有将软硬一体化生产投入到自己的企业生产模式中去的企业将会在竞争中处于劣势。这种现象不仅出现在手机生产商中间，中国的整个IT行业都存在这样的现象。即使生产商做出了很大的努力并且极力地宣传，顾客却仍然理所当然地认为国产的硬件设备没有任何的优势。电子产品中通用的硬件，因为其价格越来越透明化，其生产商的利润在不断地下降。顾客一门心思地认为硬盘、操作系统和CPU等都是国外的产品

比较好，因此单一生产硬件产品的商家在价格上没有任何优势，反而一步步进入利润逐降的怪圈。

但是和手机行业不同的是，IT行业的怪圈远不止如此。因此，软件生产商也不赚钱。顾客对软件的投资，只占其电子产品投资的极小一部分。任何一个系统都可以用光盘安装和卸载，而一个光盘的价格又非常低。因此，顾客从心理上也不愿意在软件上过多投资。由于这种原因，中国的硬件及软件生产商正面临被动的局面。

这是一种既不利于中国的硬件、软件生产商，也不利于顾客专注业务发展的信息产业格局。软件供应商失去了硬件支持，就变成了跛脚的瘸子，很难走得更远。硬件供应商不了解顾客业务需求，就成了睁眼瞎，无法把握发展的正确方向。中国的硬软件产业，一瞎一瘸，一路走来，着实不容易。即使是国内最大的软件厂商——用友，其2011年的销售收入也不过40亿元，根本无法和国际市场比较。

一体化模式的好处不仅能在IT行业体现，在企业的应用市场上体现得也非常明显。海外甲骨文公司是这个市场领域“第一个吃螃蟹的”企业。其创始人拉里 · 埃里森是乔布斯的好朋友。读过《乔布斯传》的人对他一定不陌生，《乔布斯传》就写到了他对苹果公司一体化模式的认可，并准备运用到自己的企业的市场中去。甲骨文立刻将这种模式投入到了自己的生产中。其生产的Exadata一体机，就融合了其公司自己开发的商业软件，例如数据库软件、数据仓库软件等。这款机器就是甲骨文公司的一个典型软硬件一体化产品。

继甲骨文公司之后，IBM公司也将这样的模式运用到了自己的生产中。其生产的Netezza一体机中硬件部分包括SMP主机，Snippet　Blad、磁盘仓和网络结构。软件部分集成了数据库、数据存储、数据处理及数据挖掘等软件。其中的SMP主机又由两台分别作为活动和备机的高性能Linux服务器组成。因此，Nettzza同时具备了大规模并行处理和对称处理的优点，建立了一个能极速分析PB量级数据的强大设备。

此外，Netezza通过将复杂的非SOL算法嵌入到MPP流的处理组件中，建立了一个强大的系统，这个系统能极速分析PB量级数据并以“流水线”的方式处理庞大的数据量及复杂的数据量。这种方式不仅提高了其性能，还减少了数据转移的成本。

EMC本是一家生产硬件的厂商，但是由于硬件生产的利润空间急剧下降，其竞争对手已经开始实行软硬一体化模式等原因。这家储存界的翘楚也开始了自己的软硬一体化之路。EMC在大数据方面早有布局，它于2008年收购了Smarts这家网络软件开发商来提高自己的网络管理能力。不仅如此，2011年，Greenplum这家Oracle、Netezza和Teradata等老牌厂商的挑战者企业，因其能够做到超出传统数据库软件10~100倍的性能被EMC收购。同年10月，EMC收购了Zettapoint，这是一家数据优化企业。2012年，EMC又收购了Pivotal　Labs和Watch4Net来提高自己产品的计算能力和绩效管理能力。一系列的收购之后，EMC成功地转型为软硬一体化的企业。并购了Greenplum之后，EMC开发并推出了统一分析平台，来加强自己在大数据方面主要提供存储和统一分析的能力。

小米手机对“米粉”需求的文化挖掘

小米公司的一位高管在接受记者采访的时候这样说道：“小米从创立起，业界对我们的看法经历了三个阶段。看不起——看不懂——赶不上。”的确，小米手机这匹后来居上的黑马，一夜间让老牌的手机企业都刮目相看，它的来势汹汹，出乎了所有人的意料。小米的经营理念就是以消费者为中心，以它的支持者即粉丝为中心。这样的经营思想帮助小米在其支持者的心中牢牢地占据了重要的地位，这对很多企业都有极大的参考价值。说起小米的粉丝团，最令人疑惑和关心的就是，小米是如何将其300万的粉丝凝聚在一起的？这些狂热的支持者对小米的快速发展起到了什么样的作用呢？

小米公司最早推出的手机用的操作系统是根据谷歌的安卓系统定制而来的MIUI操作系统。这个系统深受很多顾客的喜爱，甚至有人买回了安卓系统的手机之后重新刷成MIUI操作系统。小米抓住了这个好机会，不久后推出了专属于自己的小米论坛，以此聚拢了一批MIUI操作系统的铁杆用户。最令人惊讶的是，这些粉丝中的一些人，一开始只是小米的用户，然后成了狂热的粉丝，最后直接加入了小米，成了论坛的版主或者运营人员。

小米的这些铁杆粉丝都是其朋友圈里的公认的技术宅。他们的朋友手机出现问题的时候，往往都找他们来解决。这些铁杆粉丝由此成了其小圈子里的“手机专家”，因此，他们的意见几乎影响了朋友圈里的所有人。

小米论坛是这些技术粉和其他粉丝直言不讳、大展拳脚的天堂。只要粉丝在论坛里抱怨其MIUI系统哪个地方不方便、不完善，小米团队就会迅速反应。并尽快在下一个版本中修改这个问题，不仅如此，他们还会在论坛里公开表扬提出了质疑和问题的粉丝。因此，很多粉丝有了一种主人翁的感觉，这是一种神秘的参与感和心灵的愉悦感。即使只是小米的用户，他们却能够感觉到自己好像参与了小米的开发过程，甚至成了小米的系统检测者。这样，小米手机不仅仅是小米公司的专利，它真正地属于了每一个顾客和粉丝。粉丝们通过这样的参与方式，对小米产生了极大的亲切感。每个操作系统都有漏洞和问题，但是使用MIUI的乐趣就在于能够亲自参与其中，解决问题。

小米在这么短的时间内取得这么大的成功，其营销手段成了各大媒体争相报道的头条。小米的营销思想类似于白酒营销中的盘中盘思想。盘中盘营销思想就是由公关某地的显要阶层带动其他阶层的营销手段。小米正是运用这样的营销思想赢得了“米粉”们的忠诚。小米的新机发布会上，到场的营销商、合作伙伴的人数远远比不过狂热的“米粉”。

在惊叹小米拥有如此之多的忠诚、狂热的粉丝的同时，仔细分析小米对其粉丝的发掘和维持，不得不由衷地赞叹，小米发展到如此地步，的确有很多值得学习的地方。全国各地的小米粉丝组成了一个个小组织，小米的工作人员正是通过这些分散的组织更有效地和小米的铁杆粉丝沟通，并解决其中出现的问题。在小米的论坛上，不同的社区都有各种技术帖。在解决问题的同时，论坛也成了小米向顾客宣传和销售的良

好平台。小米的粉丝与其用户的重叠度非常高，为了让更多的用户转变为自己的粉丝，小米组建了专门的团队以保证社区的有效运行。

微博成了如今最流行的社交平台，小米的官方微博上粉丝数量大约300万人，巧合的是，其手机的销售量也是300万，这也非常有力地说明了小米的粉丝和其用户重叠度非常高。和论坛相同，小米同样组建了专门的团队来管理自己的官方微博，以保证和粉丝的及时沟通。

从一开始的被忽略、被质疑，到如今小米已经在其顾客心中占据了重要的位置，制造商小米和它的消费者及粉丝之间的隔阂正在逐渐缩小，他们之间的交集也越来越多。很多粉丝一路伴随着小米成长，最终成了小米的员工。没有成为小米员工的粉丝，也通过社区积极地加入了小米的研发和测试环节之中。消费者和制造商之间形成了如此亲密的关系，小米如何不壮大？这两股相互促进的力量，正是帮助和推动小米发展的最大动力。

小米和其消费者之间的关系是一种新型的买卖关系，这种良好的关系中，小米做到了以消费者为中心，以消费者的需求为标准定制化生产自己的产品，也因此获得了更广阔的顾客群。和其他企业不同的是，小米的客服不仅限于其全国统一电话，其微博、论坛都是有效的客服部门。

小米的成功使得它的这种运营模式被很多企业争相模仿，不仅要让顾客知道企业能够听到他们的声音，更要让顾客知道自己可以介入企业的各个环节。很多企业的公共平台，例如微博和官方网站都只是宣传的

摆设，不起任何实质性的作用。这也正是这些企业需要向小米学习并改正的地方。

阿里巴巴数据化运营的那些“大招”

提到数据化运营，很多企业花费了大部分的时间思考要去做什么的问题。但国内电子商务行业的龙头老大阿里巴巴却没有走这条“寻常路”，当阿里巴巴开始数据化运营的时候，他们想到的是“人”。的确，首先要从“人”做起，才能让数据化运营落地。阿里巴巴的秘密就是简单又有效的三招。

第一招，找数据。

企业要实现其经营目标，离不开数据分析，因此也就离不开能够胜任数据分析的数据分析师，这一类人最懂分析什么样的数据，如何分析。但是并不是去找一个专业理论非常丰富的数据分析师，就可以高枕无忧了。很多数据分析师，在专业领域数一数二，但是空有一肚子的理论，缺乏商业意识。他们为企业分析数据的时候，不懂得究竟要运用哪些数据去分析，于是就成了“盲人”。这种数据分析师对企业是有害的，因为他们的分析结果对企业决策层没有任何的参考价值。

很多数据师仅仅把没有整理或者初步整理的不具代表性的数据直接交给了CEO，并且他们没有向管理层解释这些数据背后的含义、体现了什么用户什么样的行为、数据的横向和纵向比较有什么结论等等。这也是导致了很多CEO每天都因为要看一大堆零碎的数据而一直抱怨的原因。

CEO需要知道的是这些数据是否精确有意义、反映了哪些问题，数据中反映了哪些市场的新现象以及需要做出什么样的决策来应对，而不是花费多余的精力来查阅资料解读数据。这就需要企业拥有一名具有商业意识的数据分析师，对商业数据有敏感的触觉。例如，看到网络上婴儿奶粉的销量忽然增高的时候，就可以预测到其他婴儿用品：婴儿推车、婴儿纸尿裤等的销量会随之上升。

数据师的商业意识并非天生的，没有任何一项技能是与生俱来的。因此，和很多其他技能一样，数据师想要拥有敏锐的商业触觉是需要锻炼的。

这个锻炼包含了很多层意思，对于数据分析师来说，他们需要多和业务部门的人接触，因为这样他们才会知道业务部门的员工每天面对的业务是什么样的。要坚持不断地进行这种锻炼，商业触觉才会逐渐增加。此外，如果数据分析师能够经常参加业务部门的周会、规划会议等活动，就能够尽快地开发自己的商业触觉。很多数据师甚至选择去业务部门轮岗，和他们一起上班、午休、喝茶聊天等，以最近距离了解业务部门，以最快地了解市场。

企业高管需要的数据分析师，是能够在每周发给CEO的周报里能看到对数据的分析，能够准确地把握市场的方向。没有任何一个CEO喜欢只有数据没有任何分析和结论的周报，这就要求数据分析师一定要有意识地和业务部门的人沟通，经常了解企业业务的情况。

第二招，沟通数据。

这一招是三招里面最重要、最关键的，并且是将前两招联系在一起的纽带。数据分析师能够从数据中看出业务的问题，或者根据业务来分析数据，这就做到了第二点。如果数据师和业务部门的员工经常联系、经常向他们了解情况，那么就会在看到数据之后，很快分析出数据背后的含义。

很多行业都非常重视数据分析及其结果，尤其是电子商务行业。随着互联网购物的普及，电子商务行业对数据的要求越来越高，也越来越依赖于数据。但遗憾的是，即使在电子商务如此发达的今天，也很少有电子商务企业能够在数据分析的环节上做到尽善尽美。很多公司的专业人员在收集数据的时候，会发现数据非常混乱，并且不同的数据杂乱地分散在很多人员和主管手中，给分类和整理造成了很大的麻烦。并且，很难将这些非常凌乱的数据联系起来并分析出其中隐含的内容。

当然，必须承认的是，在数据运营的时候，会存在很多主客观的因素，影响数据和数据分析的精准度。数据本身是没有思想的，对数据的解读在某些程度上也会受到和产品相关的各个部门的人员偏好的影响，这样，数据就有了思想和针对性。分析这样的数据，就会有不同的结果。例如，市场部门和运营部门对“产品转化率”的理解有很大的不同，如果这样的分歧一直存在，那么企业进行数据分析，其结果的波动性就会很大。

正是因为这些不稳定因素的存在，很多问题最后都要归结到“人”和“企业”这两方面来。如果数据分析师不能和其他部门的人产生

良好的沟通，不能深入地了解业务部门的每天情况的变化，那么即使再多良好的数据也只会被浪费。数据分析师必须做到以一个业务员的身份来客观地分析手中的数据，才能给管理层一个较为真实的分析结果。很多有经验的数据分析师，面对数据，能够在很短的时间内给出中肯的分析建议，这正是将第二招牢牢掌握的表现。

不得不提的是，第二招有两个具体的场景。例如，面临一堆数据和一个特定的商业场景的时候，当能够准确地把握二者之间的关系时，就表示能够实现对二者之间存在的“数据中间层”进行准确地把握。

第二招的另外一个层面，存在于企业中不同组织的数据之中。即当存在某一个特定的商业情景或问题的时候，需要对甲、乙、丙、丁四个不同组织之间的数据进行互相联系并分析，才能得出结论或解决问题。例如，某一个销售日，由于UV一下子有很大的增长，那么一定要去看退款的数量，如果退款的数量在这一天之内也有很大的增长，那么这就是一个正常的现象。当然，如果这一天的UV并没有不寻常的增加，但退款的数目却异常地增加了，那么就要考虑到可能是少数买家为之，此时，就要综合其他各方面的原因进行全面的分析。这就是将不同组织之间的数据融会贯通之后分析的必要性。

很多企业，在对这至关重要的第二招的把握上存在很大的问题。主要表现在两个方面。第一，从人的角度来说，即数据分析师的主观原因造成的数据分析的不合理。如果数据分析师对企业不同组织架构中的数据不能明确地分辨，那么不仅会造成分析结果的偏差，还会导致更加严

重的后果。然而，很多企业之中都有这样的情况，部门与部门之间沟通不畅，导致数据和数据之间关联性较小或者立足点不同，这就给分析造成了极大的困难，更别说得出有意义的结果了。

第二招，即打通数据与数据之间的联系，是在数据分析中非常重要的一点，也是能够进行准确数据分析的基础。因此，事先做好一系列铺垫工作非常重要，例如保证数据的安全性，统一不同部门统计数据的标准，以免造成更大的工作量。最后要保证不同部门数据之间能够顺利地相互交换，使得部门之间有较好的沟通和了解。第二招可谓是三招之中的核心，尤其是在现今每个企业的数据都越来越复杂，基数越来越大的时候。完全凭感觉去分析数据的时代已经过去了，面对海量数据，必须做好打通数据之间联系的工作，才能更好地分析数据，使企业做正确的有益的决策。

第三招，对数据进行运营和分享。

用好第三招的主要内容就是能够通过对数据的分析得出这样几点结论：企业的业务是否正常，如何对数据进行优化来促进企业业务的优化，如何通过对数据的分析来找到有益于企业发展的方法，帮助企业创造新的商业价值。这些问题之间看似存在明显的递进关系，但实际上却不是如此简单的逻辑关系就能够解释的。对待这三个问题，要根据不同的场景来具体分析。不同的问题有不同的解决方法，要具体问题具体对待，做到对症下药。以下就是几种能够解决不同问题的不同方法。

任何事物的发展都需要一个范围作为约束，数据也一样，需要一个

具体的框架来具体分析企业的业务水平究竟如何。因此，给数据搭建框架非常重要，有了合适的框架，才能对数据进行更加准确的分析，也就能更加直观地分析企业业务的好坏。数据的框架，就是一个标准，能够将数据在同样的层面下进行分解的标准。指标化分解是一种重要的分解方式，能够将混乱的数据整理出条理，并客观地分析企业的业务。

这样的方式就类似于生活中，因为感冒而去医院检查，医生首先要求验血来判断是不是病毒性感冒一样。根据客观的数据得出真实的结论，然后对症下药，效果才会立竿见影。看数据要能够看到数据背后隐藏的信息，而不只是被表面的信息所迷惑。对数据进行真正的分析，就会发现很多数据和其表面所代表的内容差别很大。例如，一个网站当天成交额的提高达到百分之二十，表面上这是个值得兴奋的数据。但是，具体分析才发现，销售额的增加是因为企业加大了对广告的投入，一对比发现，成交额的提升带来的收入增加不足以覆盖增加的广告费用，那么这个看似可喜的消息就立刻变成了一个可悲的消息，这说明企业的广告效果非常差。

很多电子商务企业，在评价自己的业务水平时，通常用到以下两套指标。第一套是企业用来计算其成交额的，公式为成交额=流量×转化率×单价；另一套指标多用于企业对商品进行促销的时候，公式为，即大促成交额=预热期加入购物车的商品数×商品单价×经验转化率×经验成交额占比。前者是用来评价企业的某一类商品或单个商品的健康度的，后者则是在企业促销的前提下，用来预测大概成交额的。

业务水平的好坏是通过对数据分析之后进行比较才能得出的，单个的数据分析得到的结论是不具备代表性的。不能够进行横向和纵向比较的数据是没有任何意义的。比较的实质就是找一个参照物，参照物的不同，会导致结果有很大的差别，因此进行比较不仅必要而且重要。寻找合适的比较对象是比较最重要的一个环节，例如，当企业进行促销的时候，需要和往期的促销活动销售额、促销幅度、顾客对促销的评价等方面进行比较。而不是和上一个季度的正常销售进行比较，因为毫无可比性。选错了比较对象，会导致数据分析产生极大的偏差，影响对某一决策内容的判断从而影响以后的决策方向。

当然，除了对数据框架的构造之外，能够使数据分析锦上添花的一个方式就是好的展现形式。常见的数据分析展现形式就是表格和图形。但是这两种展现形式在某些特定的情境下是不可以互换的，否则会造成分析结果的不直观和偏颇。使用了不恰当的表现形式，会使得对数据分析结果的分析产生障碍甚至误解。那么如何选择合适的数据展示形式呢？有几种便捷的选择原则：当需要对精确的数据有展示的时候，应该用表格，此时用图就显得非常不合适。好的数据展示形式有利于决策者根据数据做出更加合理的决策。

数据分析的最终目的就是通过对数据的分析，发现并解决问题。那么，这样的目标就要求数据分析要准要切中问题要害。以下，举一个例子来解释如何利用数据帮助业务为企业发现更多的商机。

想要利用数据帮助业务让企业发现并抓住更多的机会，就必须了

解，这样做的价值点在于通过这样的方式，使得企业的数据变成了人人都必须使用的数据。长此以往，企业的每个人都会变成数据分析师。这正是最理想的状态，每个人都能够对数据做出分析并给出判断和结论。客观的数据是最好的指示器，如果能够对数据进行客观敏锐的分析，那么每个员工都可以成为企业决策的参与者。

当员工积极地参与进了数据分析的工作中之后，就能够对企业出谋划策。帮助专业人员在产品的个性化设计、销量预测等方面出一份力。这样的模式也减轻了数据分析师的前期工作，使得数据分析师能够对数据做出更加精确的分析。如此一来，数据的运营就进入了一个良性循环，有助于企业快速地提升自己的业务水平。

在阿里巴巴流传着这样一句话："让信用变成财富"，的确，阿里巴巴也在践行这句话。阿里巴巴所有运营程序的核心就是通过数据来计算客户的信用水平，并且通过对客户信用的评估来为客户进行授信。一系列流程之后，通过审核的顾客就会获取他所需要的资金。正是这样一个良好的信用模式，使得阿里巴巴越来越成功，旗下的产品也越来越多，深受顾客的喜爱。

数据分析师工作的一个重要环节就是通过对数据的分析和对市场的预测，说服产品经理某样东西可以被列为产品出售。但是这样一个看似简单的环节在实际操作中却很难顺利地实行。尤其是公司对于新产品的开发和上架没有任何问题，但是如果数据的分析达不到这个效果，一切都是空谈。

因此，如果不能有效地获取、使用、分享、协同、连接数据，数据化运营就会受到很大的阻碍。于是，数据分析师是否能将数据简化，留下最重要的部分就变得尤为重要了。千万不能出现收集了很多的数据不知道怎么用，不知道用的数据是怎么来的等情况。这就要求能够熟练掌握前文提到的数据化运营的三招，让每个员工都成为数据分析师。

大数据中的企业价值及客户价值

有效的数据收集不仅能够实现企业资源的合理配置，也能够更好地服务顾客，为顾客带来更好的体验。如何有效地收集数据呢？数据的收集包括两个维度，第一个维度，是衡量数据对企业的生产价值，第二个维护是衡量数据对顾客的价值。第一个维度要求记录顾客行为数据里对企业有价值的那些。第二个维度则能很好地帮助企业为顾客提供更好的服务，创造更大的顾客价值。

从第一个维度企业的价值来看，数据收集的主要目的就是实现了对企业资源的合理分配。可以通过不断的升级自己的操作系统，让自己的产品能够更方便地被顾客找到，从而增加产品的二次甚至多次销售，创造更高的销售利润。

从第二个维度来看，企业的目标就是要提升顾客体验。例如，提供各方面的服务界面和搜索引擎，方便顾客的购物，增加顾客消费的喜悦感和满意感。

这两个维度看似相似，实际不同。从不同的维度看数据，也会有不同的结果。例如，在工作日里，银行门口总是排着长长的队伍。从第一

个维度去看，则需要知道每天固定时段排队的人数，然后设法减少顾客排队等候的时间。从第二个维度去看，则要思考这些排队的顾客平均要等待多久才能获得相应的服务。同样的问题，从不同的角度去思考和理解，就会得到不同的结论。两个维度所看重的信息点不同，第一个维度的解决方法可以是开发手机客户端，减少排队等候的人数。第二个维度的解决重点就是减少顾客等候的时间，例如在高峰期多开几个窗口等。

因此，在分析数据的时候，要分清是从哪个维度去解决问题。因此，在大数据时代，数据运营要衡量不同的方面。同时，还要注意看待数据的不同角色，他们的感受和体验也是不同的。即在银行看来，20个人的队伍并不算多，但是在顾客心中，等待超过了10分钟，就属于长时间的等待了。

排队只是一个典型的例子，电子商务企业在操作系统方面的实践和排队相似。企业思考的是如何让顾客更多地点击自己的商品，而消费者想的却是："我想要买的这个东西能否顺利买到呢？"

因此，企业在对数据进行了分析之后，只用通过改善自己的推荐系统来更好地改善用户的购物体验。此外，对用户的整体购买情况要做横纵向的分析，形成一条完整的数据链。例如，如果企业想要了解用户登录前的浏览情况，就需要更新系统来达到这个目的。

第四章 数据和企业管理，高层更看重大数据

大数据带来的好处让很多占得先机的企业尝尽了甜头，其后的其他企业相继模仿、穷追不舍。要在这个充满机遇的时代打下一片江山，需要企业上上下下都对大数据有敏锐的把握，专业的综合人才是能够带给企业全新的驱动力的主力军。

沃尔玛如何用数据构建管理模式

家乐福和沃尔玛等大型超市，已经成了所有市民生活中必须去的地方。这些大型超市进入中国市场之后，中国市场的零售行业发生了很大的变化。沃尔玛的优质服务和高销售额，使得很多企业都争相研究其管理方面的方法和经验。例如，沃尔玛要求所有的工作人员都要学会微笑，见到顾客的时候保持标准的微笑，并且对员工进行各方面的培训。但是，很多人不知道的是，沃尔玛的管理经验的核心之一就是它科学化的数字管理，沃尔玛也需要利用对数据的分析来解决销售过程中的问题，提高顾客的满意度。

顾客在逛超市的时候，不会只买一种商品，而是会买很多相关的商品。例如，顾客想要购买大米的时候，就会顺带购买做菜需要的蔬菜、油、各种调味品等。顾客在购买洗衣粉的时候，就会想到购买卫生纸、香皂等日用品。很多顾客的购买动机都是偶然的，很可能因为一个降价的标志，而购买了很多原本没有想要购买的商品。沃尔玛将大量的数据整合分析之后，发现一条规律：如果商品之间具有一定的相关性，一般为互补品关系，就会增加商品的销售量。例如，沃尔玛通过数据分析发现，超市里蔬菜、肉类和食用油的销售比例为100：80：10。足以证明上述的规律是普遍存在的。

职业经理人经常会来到销售现场，做一些检查巡视的工作，但是除此之外，他们同样非常关注其他相关数据的变化。在一些大型的连锁超市里，收银台会随时收集顾客的采购信息，并将之传送到后台的ERP信

息系统，从而进行统计。有的超市7点才开门，9点就已经做好了数据汇总的工作。这样一来，经理就可以通过汇总报表观察超市各商品的比例，以蔬菜、肉类和食用油为例，这三种商品的正常比例应该为100：80：10，但如果数据显示结果为100：40：10，就说明肉类的销售出现了异常。销售了100单位的蔬菜的同时，本应销售掉80单位的肉类，但是今天肉类的销售却降了一半，经理会立即跑到肉类销售区去查看原因。看是价格的、陈列的，还是质量的问题。一旦发现原因，就可以立即进行有针对性的调整。这样，问题的苗头刚出现，就会被迅速控制和改善。这就是“即时干预”管理法。

如果没有这些数据，超市经理的观察感受就是：超市的员工非常忙碌，货架前站着络绎不绝的顾客，一切看起来都很顺利。超市经理在巡视中能发现的问题，大概也就是类似商品摆放不当，员工服装不整齐之类的问题，而最关键的要点则很容易被忽略掉。

在对每天的报表数据进行统计和分析后，经理可以发现更多的问题。比如，一周后的某一天，超市7点开门后，经理9点看到蔬菜、肉类和食用油的销售比例是50：40：5。商品的销售比例没有问题，但销量却出现了问题，这是因为整体客流量降低了。这时候就要去查明，为什么今天的客流量整体减少了？

“相信数据，用数据说话”，在沃尔玛等国际型企业中，已经成为职业经理人的思维惯式。

让大数据进入企业管理

人的经验和感觉经常会与事实有所偏差，而在大数据时代，无论做什么工作，管理者都需要学会用数据做管理，而不是凭感觉。

随着不断地发展壮大，很多中国中小企业都开始进入大型企业的行列。随着中国企业对规模化管理的需求的加大，很多企业都迫切需要解决很多一直存在的问题。比如，复制和扩大规模，提升执行力，建立企业文化等等。而几乎所有这些问题，其实都可以通过数量化思维解决。根据企业运营产生的各项数据，管理者可以清楚地看清现状，并做出正确的决策，从而让企业处于良性发展之中。

关于管理定义的描述多到让人眼花缭乱，但是无论是哪种定义，都离不开“目标”这两个字，管理其实就是相对于目标而言的。那么，目标到底是什么呢？

（1）将数据作为方向

以数量化思维来理解，目标其实就是“目的”的数量化标准，而“目的”则可以理解为企业愿景或个人梦想，比如成为全球领先或著名学者等。

目标则是“目的”的标尺，是对于“目的”的范围和程度的数据定义。比如，你的目的是成为全球领先，什么是全球领先呢？需要100亿美元的销售额，并在同行中排名全球前五；比如你的目的是成为著名学者，什么是著名学者呢？得有10本专著、100篇刊登在核心期刊上的论文，并在国内多所知名大学任教。

目标是员工奋斗的方向。如果管理者在下达指示说：“把这项工作做

好”，员工只会感到迷茫，他们不知道要做到怎样的程度，才叫作“做好”。不是每个员工都能够领会和把握管理者的要求，在不断的揣测中，误差在所难免。

在目标制定过程中，有一个重要的原则——“SMART”原则，而其中最为关键的就是“M（measurable）”——可度量，即将绩效指标数量化或行为化，并确保相关数据的可收集。M往往是中国企业管理的弱项，管理者最喜欢说的就是“尽快”“尽力”……而对于“尽快”和“尽力”的理解，员工和管理者往往有所偏差。

如果用数量化思维来做，结果就大不相同。比如管理者想要给老员工安排深化培训，自然不能简单地说“为老员工安排进一步的培训”，而应该下达这样的指示：

目标人群：在公司工作五年以上的员工；

培训时间：2个月内；

培训内容：与初期培训内容有所差别，添加沟通技巧、时间管理方面的培训内容；

培训结果：调查评估达到80分；

负责人：培训部门主管

管理者在处理企业管理工作时，要避免使用不清晰的词汇，而应该以可度量的数据作为员工努力、公司发展的方向。

（2）用数据复制成功

很多人喜欢到肯德基、麦当劳这样的餐厅就餐，因为无论在哪个城

市的哪家分店，它们的价格、服务质量、餐饮质量都相差无几，它们是怎么做到的呢？通过尽可能地以自动化设备烹制，来确定各种食品的加工时间，并详细制定了各环节员工的操作步骤，将其精确到秒，肯德基、麦当劳就有了一套可以复制的标准流程，这就是定量规范的原则。

“复制成功”，就是将局部的成功经验进行归纳总结，将其标准化、流程化，以精确的数字进行描述，从而直接复制到整体中去。而这样的成功方式，离不开数据作为依据和标准。

中国餐饮很难做到麦当劳、肯德基这样的全球连锁，就在于中餐食谱无法复制。中餐菜谱中经常会出现“几块”“几条”“捶松”“沾满”“油六成熟”这样的词。可是，一块到底是多大？一条到底是多长？捶松是松到什么地步？沾满是沾到多厚？六成熟是多少摄氏度？很多人不知道，也因此，同样一个食谱，每个人做出来的味道却都不一样。

在中国的文化背景下，感情和道德被很多管理者作为管理标准，并将之作为决策依据。数量化管理与中国这样的文化背景相比，可以说是格格不入，这也是工业革命后，西方国家得以超越中国的原因之一。

（3）数据提高执行力

越来越多的管理者开始关注企业执行力的问题，这是因为企业员工在相互推诿中，经常会“说了不做，做了也做不好”。

中国有句古话，“没有功劳、也有苦劳”。很多管理者在考核员工时同样抱着这样的心态，然而，企业经营就需要盈利，有的企业想要成为全球领先，有的企业可能只想着维持生存，但无论如何，都需要盈利作

为结果。如果在企业管理中，忽视了结果，只以过程而论，那即使员工超负荷劳动，最终也可能只是“徒劳无功”。

很多中国企业在执行力上都有所缺陷，归根结底就是因为缺乏数量化的管理理念。在数量化的管理理念下，决策或者计划的执行就有章可循，当一切以数据说话时，员工也会只看自己的业绩结果，而不是自己付出了多少劳动。

职业经理人需要明白，自己最终追求的是一个量化的结果，而不是员工的工作劳动，因此，职业经理人的工作必须以结果为导向。在工作中，必然存在各种各样的风险，让员工的劳动化为乌有，但这并不能成为借口。在管理者的职业发展和收入提升中，量化结果才是关键。优秀的职业经理人会以完善的准备工作，规避风险，并积极地采取行动，以一种坚强的态度面对恶劣的环境，而不是互相推诿或找借口。

什么是以结果为导向？首先，得以风险预估和细节管理，来规避风险、提高效率；其次，在工作中，不要找借口，而是以结果为衡量工作的唯一标准。西点军校最重要的一条行为准则就是——“没有任何借口”，每个人都应该想尽办法去完成工作，而不是想尽办法为失败找看似合理的借口。完美的执行力，内涵就是敬业、责任、服从和诚实。

职业乞丐脑袋里的大数据

无论做什么工作，都需要良好的职业精神。什么是良好的职业精神呢？同样要落实到数据上，要用数据去分析和解决问题。网上流传着这样一个关于乞丐的笑话：

有一天，小张刚逛完商场，买了一条名牌牛仔裤，他在商场门口等一个朋友。这时候，一个乞丐直接来到小张的面前停住。不出意外的，这个乞丐开口就说道：“先生，行行好，可怜可怜我吧。”小张也懒得躲开，就扔了一个硬币给他，闲得无聊的小张，还和乞丐聊了起来。

乞丐倒也不赶着做下单生意，而是说道：“我一直在这条路附近乞讨，一看到你，我就知道我要有收入了。能买Levi’s的主，都不会太吝啬的。”小张很是惊讶：“懂得还挺多啊！”

乞丐说道：“乞丐也不是那么好做的，也得有方法。”小张一愣之下，也十分好奇：“什么方法？”

“我做过计算的，这里是市中心，每天这里的人流量都超过1万人，虽然很多是穷人，但更多的还是有钱人。按理来说，只要我每天都能跟每个人讨到一块钱，我一个月就能挣到30万。当然，不是每个人都会给钱的，我也没办法一天向一万个人伸手。那我就得好好利用我的时间，找到能给我钱的人，才能挣到更多的钱啊。”小张想想确实如此，乞丐又继续说道：“我乞讨了这么久，也计算过，在这个商场门口，每天的人流量中，有三成是我的目标客户，有70%的概率会给我钱；而两成是潜在客户，成功概率是50%；至于剩下的五成，我可没时间浪费在他们身上。”

小张又问道：“你怎么看待你的客户的？”

“首先是目标客户，这又分为两种：第一种就像您这样的年轻先生，有钱又大方；第二种是那些情侣，男人怕丢面子，就会给钱的。其次是

潜在客户了，就是那些一个人的漂亮姑娘，她们总是不想被乞丐纠缠，所以会‘破财免灾’。而无论是哪种客户，我都会选择年龄在20-30岁之间的。太年轻，没钱；年纪大，有钱，但男人的钱在老婆那里，女人又抠门。所以呢，找他们要钱，那可就太难了。”

“那你一天能挣到多少钱？”

“工作日的时候差点儿，就三百左右吧，但是周末的时候，能有六七百。”

“这么多？”小张不禁有点怀疑。

乞丐就又给小张算了笔账：“我也不会让自己太辛苦，和你们一样，我一天就工作八个小时，从中午11点到晚上七点，周末肯定是要出来的，但周一到周五的时间，我会休息两天。我每次伸手呢，大概是5秒钟的时间，当然，还有来回走动和搜索目标的时间，所以大概1分钟能乞讨一次，8个小时就是480次。再乘以一下概率——（70%+50%）/2——就是60%，就是288块了。”

小张越听越感兴趣，觉着这人做乞丐真是屈才了，完全可以胜任一个市场营销总监了，就让他继续说。

“很多人觉得乞丐其实是靠运气吃饭的，可不是这样的。我就这么说吧，看到一个小伙和一美女一起出来，你选哪个乞讨？”小张想了想，说不知道。

“应该去找小伙，身边就是美女，他好意思不给吗？你要去找女的啊，小伙直接走过去了，美女也装着害怕远远躲开了。”

小张还在咂摸着其中的道理，乞丐就说了："所以说啊，得职业一点儿才能挣钱啊。"小张听过十几个老总说过这句话，从来没想过会从乞丐的嘴里听到。"大多数乞丐一个月就讨个两三千块，我呢，每个月都是将近两万……"乞丐还没说完，小张就激动地说道："你有兴趣收我做徒弟吗？"

虽然只是个笑话，但事实确实如此，无论是什么职业，即使是一个乞丐，也能依靠职业化挣大钱。

职业经理人与大数据

既然乞丐都能依靠职业化挣大钱，那么真正的职业经理人应该如何去做呢？刘峰在一家国际著名的通信企业工作，经过八年的工作，他已经晋升为了公司的产品部门经理，负责新款手机在中国的产品定型和推广工作，看看刘峰是如何来做的吧：

（1）计划是第一位。

每天早上8点，刘峰就会从家出发，开车前往公司，而公司的上班时间是9点。刘峰计算过，家里距离公司10.8公里，8点出门的话，就会避开早高峰，只需要30分钟就能到达公司。但如果踩着9点去公司的话，就会融入到早高峰的车流中，得45分钟才能到达公司，也就是说得8点一刻出门。刘峰就想了，与其时间浪费在堵车上，还不如提前15分钟出门，还能多出来30分钟为一天的工作做点准备。

而在去公司的路上，刘峰还会把当天要做的工作都拿出来想想，看看哪些是重要、紧急的工作需要优先处理的，哪些是可以授权给下属去

做的，哪些工作不那么紧急。

今天，刘峰在路上把自己的工作大致梳理了下，包括：向领导汇报新产品的开发进度，检查督促下属的工作进度，与供应商讨论新产品的生产，另外，还要再招3名工程师，研发后两个季度计划推出的3款手机，还有一些突发的临时性工作得做好准备。刘峰想了想，还是招聘的事情最重要，毕竟，这关系到公司的下一步发展。

8点半的时候，刘峰到达办公室，他的第一件事就是打开电脑查收邮件。因为公司总部与北京有6个小时的时差，所以刘峰经常会收到总部同事夜里发来的邮件。今天早上，刘峰打开邮箱时，里面已经有了18封邮件。关于查看邮件，刘峰也有着自己的方法：刘峰首先会查看那些只发给他一个人的邮件，因为那里面通常有关于他直接负责的事务；接下来，刘峰才会查看群发的邮件；最后，则是那些抄送的邮件。在快速查看邮件的过程中，刘峰会对一些简单的邮件，直接回复过去，重要的则标记好，而不太重要的则直接转发给下属。当刘峰处理完今天的18封邮件时，已经是8点50分了。

接下来，刘峰就开始思考关于招聘的工作了。现在，想要招几个既专业又懂市场的人才，实在是太难了。虽然猎头招聘能找到这样的优质人才，但因为部门的预算有限，手头的经费只能支持他通过猎头招一个人才，另外两个则只能想其他办法了——拓宽招聘渠道成了当务之急。

（2）工作安排要周密。

8点55分的时候，秘书小董来到了公司，两人打了个招呼，刘峰就让

她休息一会，等十分钟之后讨论工作安排。

刘峰想要小董负责招聘工作，但小董既没有这方面的背景，也不懂得关于研发的专业知识，因此，刘峰必须得详细地教导小董怎么去做才行。

十分钟之后，小董来到了办公室，刘峰就详细说了关于这次招聘的工作：首先，刘峰介绍了这次招聘工作的背景、困难和重要性，以及关于所需人才的具体指标；然后，刘峰明确告诉小董，希望她在目前4个主要的行业网站上，以及微博上发布招聘信息，并提醒她注意，行业网站的招聘内容与微博上的要有所区别，至于具体如何去做，可以自己“百度一下”；刘峰又询问了小董今天既定的工作计划，经过一番计较之后，刘峰就提出了要求，让小董在下午2点之前完成草稿发给他，等到审核通过后，要在下午2点之前发布完成，并反馈各网页招聘信息的截图；最后，刘峰询问小董是否能够做到，小董自然给予了肯定的答复。

（3）量化分析很重要。

安排完招聘工作，已经9点半了，刘峰来到了会议室——今天有一场各部门经理都须参加的市场会议，会议的主要内容其实就是关于马上就要到来的“小长假”的促销事宜。市场总监直接就问道：“促销目标大家都知道了，那应该采取怎样的促销形式，要投入多少预算呢？”

刘峰其实早早就做好了准备，于是直接递交了自己的分析数据：根据近五年的销售数据，在对信函直邮、媒体广告、网站抽奖等推广形式

的每年投入产出比进行分析后，刘峰还根据市场环境，指出了每种推广形式在未来一年的升降趋势。最后，刘峰提出应该将推广重点放在“网站抽奖”上，因为近几年来，这种推广形式的参与人数平均每年上升达到38%，而投入产出比则平均上升29%，而其他推广形式则表现平平。

接下来，大家重点讨论了如何去做网站抽奖，而市场总监也对这种方案表示支持。

（4）决策需要集思广益。

10点半的时候，公司资深调查顾问陈杰如约到访，在一个月之前，刘峰就将关于新产品的20份设计稿交给了陈杰，让他的团队做好市场调研工作，并从中选出三款最受欢迎的界面。由于新产品的目标客户是年轻白领，刘峰曾经找到200个不同年龄和消费水平的白领进行调研，从而在500多个设计稿中选出了这20个，但因为自己精力有限，因此，刘峰需要找专业的调查公司来做更细致的调研工作。

陈杰的调查范围涉及2000人，并按照有无购买过本品牌产品以及年龄、性别等进行了分类调查，最终选择了三款大家比较喜欢的设计，从而保证新产品上市后能够尽快得到目标客户的认可。

（5）以结果为导向。

11点多的时候，刘峰突然接到研发部门的电话，说是遇到了一个难题：因为原定电池供应商无法购进足够原材料，因此，无法按时提供原计划的手机电池，而替代电池在电量上又有所缺陷，将会影响到新产品的客户体验，可能会使新产品的销量面临困境。

刘峰并不负责新产品的研发，但作为公司今年的旗舰产品，如果该产品的销量出现问题的话，今年的销售任务也将无法达成。到那时候，领导不会区分谁是研发负责人、是不是电池供应商的问题，而是会直接找刘峰这个产品部门经理。因此，这个问题出现后，刘峰第一个想法就是如何解决，如何保证结果。

幸运的是，刘峰曾经在研发部工作过一段时间，有着自己的一些人脉，于是，他立刻致电某电池领域专家，询问解决方案，并让自己的得力副手立刻赶往深圳，与当地的研发部门一起找到替代方案。

想要成为像刘峰这样的职业经理人，就必须在工作中运用数量化的思维方式和工作方法，如果老板想要拥有这样的人才作为下属，更需要将数量化融入到企业文化之中。

企业组织管理不介入大数据，就要被淘汰

传统企业组织的发展必然会经历这样几个环节：初创期，只是一个几个人的小团队，大家相互帮助，每个人都需要负责几种工作，分工也很不明确；随着业绩的增长、规模的扩大，企业首先会实现专业化的分工，确定各部门、各岗位的职能，并逐渐形成层级，提高管理效率，确保企业决策的上通下达；当企业进一步拓展，业务种类增加、经营地区增加时，总公司领导分公司的组织结构则会使工作效率降低，事业部制随之出现；当企业发展到相当的规模时，则会采用矩阵制……而随着企业员工的增多，企业的沟通效率也逐渐降低，与此同时，管理人员也会出现冗余——这就是所谓的“大企业病”。

而在大数据时代，“大企业病”也找到了合适的解决方案。随着各种数据的收集、利用，信息得以在企业内部有序传播，从而提高企业内部的沟通效率，另外，当企业拥有一套完善的数据系统时，很多需要用运算解决的决策工作，也可以直接依靠数据系统解决，相关的管理人员则可以进行裁撤。

（1）“合久必分，分久必合”。

在传统的企业组织结构中，为了提高工作效率，企业会不断细化分工，出现很多部门和岗位，而在这种情况下，单个的部门或岗位则很难看清楚企业整体的发展目标。而在大数据时代下，企业内部信息得以迅速、有效传播，沟通效率的提升，使得企业的每个人都能够及时了解企业各方面的信息，从而在相互协作中，实现企业整体的成功。虽然企业的部门和层级仍然可能存在，但大数据会让企业各部门、层级跨越界限，深度合作。

（2）垂直边界得以重组。

在大数据时代下，传统企业组织结构的垂直管理会愈趋扁平化。随着互联网与信息技术的发展和普及，企业在引入OA系统、CRM系统、ERP系统等企业管理系统后，其工作效率和沟通效率都得到了极大的提升，但仍然面临着很多问题，比如不同系统中的多种数据的整合、数据获取的及时性以及系统数据分析的自动化等。

而这些都需要大数据技术来解决，数据将成为企业的一项重要资产，而支持结构化和非结构化数据的数据中心，则成为大企业发展的必

备要素；小企业则可以通过合作，使用一些公共的数据中心。这样一来，企业数据则可以在多个系统中迅速整合到数据中心，并推送给有需要的部门或岗位。另外，数据中心也可以利用大数据的算法，自动分析数据，从而更为迅速地形成更准确的分析报告。

随着信息的扁平化，企业的业务流程也将愈趋扁平化。立足于大数据，企业业务的中间环节将被大幅度压缩，业务流程的效率也得到巨大提升。由于在大数据管理下，企业的沟通环节更少，信息弱化自然也就越少。而且业务的支撑工作也可以预先进行处理，并实时获得营销结果，整个业务流程都得以缩短。在扁平化的业务流程中，企业也将真正地实现“以客户为中心”。

业务流程的扁平化，则需要组织的扁平化给予支持。在传统的业务流程中，各层级智能都被限定，而这与扁平化的业务流程则有所出入，会影响新流程的业务开展。因此，当企业运用大数据进行企业管理时，领导就应当用于改革，删减企业冗余部门和岗位。扁平化的组织结构将不再以管理控制为主要职能，而是以支撑服务为主。

（3）水平边界得以融合。

在传统企业组织结构中，由于各部门职能不同，其产生的数据格式也有所区别，比如市场部门的信息多是结构化和多媒体数据，而客服则多为文本数据，研发的数据则更为多样……数据的格式区别，也直接使得各部门之间的相互了解出现障碍，“以邻为壑”的现象时有发生。企业内部沟通和协作成本过大，也会使得企业无法对市场进行反应。

这一问题其实由来已久，企业家和管理学家也提出了各种解决方案，但都存在这样那样的问题。而大数据则让企业的水平边界得以融合，由于大数据时代下，各部门间信息透明化，信息也得以及时传播。当数据中心建立之后，各部门数据可以有效存储并解读，使得企业各部门终于能够形成一个整体，共同组成一个完整、全面的信息流。

在传统的企业组织结构中，各部门间的信息传递是单向的，每个部门都会出于本部门的利益对所搜集到的信息进行筛选，线性的信息传递也使得所需时间跨度变得相当长。为了解决这一问题，有的企业建立了独立的分析团队，为整个组织而不是单个部门提供数据分析服务，但分析团队与业务的脱离，也让其分析结果无法“落地”，而浮于表面。大数据则可以让数据全方面、实时地产生并相互关联，各部门之间就像电脑一样被连接在一起，从而能够实时协同工作，而信息透明化也让各信息节点之间能够相互监督，各部门的工作会自动为整个组织提供服务，而各部门间的协作效率也能够急剧提升。

在水平边界的不断融合过程中，企业也需要进行部门重组，这同样需要大数据的支持。在以客户为中心的价值链中，灵活、实时和动态协同成了企业管理的关键。而按专业分工的部门形式则会影响到组织运营效率，因此，企业需要将各部门进行重组，最终形成一种网状的组织结构，让各个节点在充分的信息支撑下，适应各种流动的非结构化数据。

（4）组织结构被打散。

在大数据资源和技术的支撑下，小团队将散发出更为耀眼的光芒。

事实上，如腾讯、谷歌等互联网企业都已经开始采用小团队的管理模式，以产品经理为核心的产品团队，将给企业带来更多的惊喜。

腾讯的微信仅用了一年半的时间，就突破了1亿的用户规模，而这样优秀的产品正是由一个小小的微信团队研发的。当然，微信的成功同样离不开大数据的支持，如果没有QQ积累下来的庞大客户群，微信即使能够获得成功，也不会发展得这么迅速。正是因为腾讯庞大的客户基础以及客户关系的数据，微信才能够利用“QQ号登录”、“添加QQ好友”的形式，迅速在QQ用户中传播开来。

谷歌采用的也是这种模式，在小团队的工作形式下，工作氛围更为良好，员工也能够实现更为高效的工作，并进行创新。谷歌几乎每个产品都是由一个小团队研发的，在小团队中，交流将变得更为有效和简单。谷歌前任CEO施密特就坦言：“小团队管理方式主要有三个好处：一是它能够增加尝试的可能性，让我们不断尝试尽量多的新生事物，这样我们成功的概率就比较大；二是能够给我们的员工更多的主人翁责任感，让他们觉得不是在一家大企业工作，在开发过程中让他们觉得自己拥有决定方向的自主权，同时又可以为用户来服务；三是能够降低团队内部协调的成本。”而作为全球第一的搜索引擎，对于谷歌小团队模式的成功，提供大数据的谷歌搜索自然也功不可没。

（5）授权更加充分、有序。

很多企业管理者都在头疼授权的问题，他们都害怕“一收就死，一放就散”：授权，害怕企业变成一团散沙；不授权，企业又无法及时应对

多变的市场。而大数据则解决了这一问题，在大数据时代，企业再不是授权给管理者，而是授权给一线员工——那些可以直接接触客户的人。企业的决策体系将从由上而下变为由下而上，决策不再依靠精英人才的管理经验，而是依靠市场数据。

海尔就开创性地采用了“人单合一”模式，重组了企业的组织结构。在“人单合一”模式中，“人”就是员工，而“单”则包括了订单、市场目标、用户等多层含义。用户是企业发展成功与否的关键要素，而一线员工在与客户的直接接触中，更是成了企业的信息终端，企业依靠一线员工了解、创造客户需求，当企业收集到这些信息之后，再以足够的资源去支撑一线员工满足客户的需求。以前是用户影响企业、企业影响员工，而现在则是客户影响员工，员工影响企业，之所以能够实现这样的突破，透明、高效、全面的数据信息起到了关键性的作用。

海尔原来的组织结构，是一种“国-省-市-县”的垂直模式，而如今则以扁平化的销售模式，直接关注县级客户。全国有2800个县，而海尔则有2000多个一级经营体，直接对这些县级客户负责。每个经营体在面向客户需求中，采用的都是“自组织、自驱动、自创新、自运转”模式，一线员工的活力和创新力都得到了极大的提升。而且，现在一线员工处于海尔销售模式的最上层，由他们去发现和创造需求，而领导则从指挥者变成了支持者，为员工服务，从而为客户服务。在互联网时代，营销模式愈趋碎片化、市场需求愈趋个性化，而海尔的企业模式则很好地适应了这一特点。

在这样的企业模式下，企业的发展节奏将自动加快。以往，企业需要由上而下地推动这种进程，而在客户需求的牵引下，企业中的每个人都会在不断获取客户需求信息中，不断改善产品和服务。大数据，让企业能够放心地进行充分、有序的授权，从而在不断适应客户需求的过程中，快速成长，企业家只需要作为一个引导者，就能让企业走上自我进化、自我超越、自我发展的健康道路。

第五章 生活无处不数据，大数据真的能算命？

大数据不是未来的某个概念，而是一步步出现在人们生活的每一个角落中，它已经对人们的生活开始产生重大的影响。敏感的企业、政府正在着手为统计、分析海量的数据，认识数据产生的统计分析结果而转变自己的工作方式方法。

未来的先兆——大数据

凡事发生前都会有先兆。现实生活当中，许多事情由于无法实时记录，看上去就好像是“人似秋鸿有来信，事如春梦了无痕”那样。互联网在实时记录方面则是“处处痕迹处处痕”。如果上网买东西，一定会先浏览、对比，再询价，如果上网搞活动的话，一定要经过征集、讨论和策划的整个过程。互联网通过服务器实现了“请求”+“响应”机制，人们大量带有先兆性质的行为数据被保留了下来。搜集完这些数据要进一步进行分析，而这些大量数据后面隐藏着的因果关系也就会被计算出来了。因此这当中的规律或被神秘化，或被庸俗化。

一切事情一定是在蛛丝马迹的先兆出现后才会发生。试想人们买一只股票之前一定会先关注一下它的走势；人们要买一样东西，也一定要先去询问一下商品的价格；人们很难会没有事先沟通而聚在一起；倾盆大雨下下来之前一定是非常闷热的天。很多书籍和文章更是大肆渲染地震前的各种奇怪的先兆。

倘若这些先兆都能被一种技术全部记录下来的话，那么显然人们都可以成为预言家。这种技术其实就是大数据技术，它们能够收集各种类型的数据，并基于数据进行分析统计，还能预测未来。和一般的信息技术相比，大数据影响的范围更广，程度更深。

IBM公司曾有一则广告，它说道：“曾经我觉得追捕罪犯是我的工作，可是现在我重新认识了我的工作。通过分析犯罪数据来识别犯罪的模式，并由此来部署警力。这样一来美国大部分城市的重大犯罪率可以

降低30%。案发之前，终结犯罪。”显然它是在利用大数据的宣传企图构筑智慧的地球。

在阿里巴巴平台的网站数据中发现，很多买家在购买商品之前会在众多供应商中货比三家，特别是查询点击的数量和购买的指数两者之间会总结出一个相对的比例，例如“2008年初，阿里巴巴网站上买家的询问量锐减，欧美的中国采购数量也随之下滑。我们比海关提前半年就已经从买家询问的数据中推算出来了世界贸易的变化，海关则必须卖了货才能得到数据。”历史上买家、卖家询问和成交的数据，统计后就能形成询问、成交的指数，两者之间有着密切的相关。作为先兆性的数据询问指数，有了活跃的询问指数才会出现巨大的成交量。马云开始发现询问指数有了异常的下滑之后，就自然而然地能预测成交量一定会萎缩。要是没有大数据的统计和分析的话，这种推测是很难做出的。马云因为向中小制造商提前预告了这一趋势而获得了非常高的声誉。

2012年6月28日，中国建设银行推出了自己的电子商务金融平台——“善融商务”。建设银行对此项业务的宣传语是：“顺应电子商务发展潮流，建设银行推出善融商务，它结合了传统金融服务优势和新兴电子商务服务的优势，从而搭建了最具综合性，涵盖全流程的电子商务服务平台。”建行内部花了很大的气力推广这项业务，尤其是分行，它们甚至是牺牲成本也要将中小商家引入平台。尽管银行推出电子商务平台，看似并非其主业，但实际上这当中正是醉翁之意不在酒。银行通过交易平台来收集各类中小商家的经营数据，推测其还款能力和贷款需

求，这样就能避免小额贷款的风险。无论建行这一次是否能成功，都说明银行的高层已经意识到大数据的重要性以及其作为先兆对于结果的预测。建行已经认识到了大数据的统计分析可以为其带来低风险和高收益，这几乎是每一家金融机构的梦想。俗话说：富贵险中求，以往必须是高风险才会带来高收益。但是当大数据的时代来临，低风险也能带来高收益，这也难怪有那么多金融机构趋之若鹜了。金融机构如果还不重视大数据背后的巨大价值的话，那么在21世纪的今天就很难在激烈的竞争中有自己的一片立足之地。

截止到2011年，互联网全球用户量已经超过了20亿。2005年RFID标签保有量仅为13亿个，这一数字到2010年就已经增加到了300亿个。从2003年到2006年，资本市场的数据增长了17.5倍。新浪微博现在每天发布的微博数量已经超过了1亿条，Facebook每天处理的数据大概为10TB，世界气象中心所积累的Web数据约为220TB，还有其他类型的数据也有8PB，等等。

国际数据公司（IDC）的《数据宇宙》报告指出，2008年全世界的数据量只有0.5ZB，到了2010年就增长到1.2ZB，从此人类走进了ZB时代。更可怕的是，从现在开始到2020年，全球数据量的年增长量都将保持在40%左右，也就是说大概每两年数据量就会翻一番。IT界此前有一个摩尔定律众人皆知，这个“大数据爆炸定律”与之极为相似。预计到了2015年全球的数据量总数要达到7.92ZB，2020年会超过35ZB，那时候的总量会是2008年的70倍，2011年的29倍。

互联网数据中心发布的《中国互联网市场洞见：互联网大数据技术创新研究2012》报告指出，到2011年底，中国国内的互联网行业已经持有1.9EB的数据总量，而这一数据到2015年将达到8.29EB以上。

人类社会的数据量随时间流逝不断地在刷新数据量级单位，从最初的TB、PB、一直到现在的EB、ZB。那么35ZB和8.29EB到底意味着什么呢？这样，先来看看数量量级单位之间的进制吧。

1B=8 bit

1KB=1024B≈1 000 byte

1MB=1024 KB≈1 000 000 byte

1GB=1024 MB≈1 000 000 000 byte

1TB=1024 GB≈1 000 000 000 000 byte

1PB=1024 TB≈1 000 000 000 000 000 byte

1EB=1024 PB≈1 000 000 000 000 000 000 byte

1ZB=1024 EB≈1 000 000 000 000 000 000 000 byte

1YB=1024 ZB≈1 000 000 000 000 000 000 000 000 byte

《红楼梦》整本书含标点的话共有87万字，依照数据计算方式的话，一个汉字等于是2B，这样的话，1EB就大概是6626亿部《红楼梦》的数据量。美国国会图书馆是全球最重要的一个图书馆，也是美国四个官方图书馆之一。一直到2011年4月，该图书馆共有1.5亿册藏书量，大致数据量为235TB，那也就是说1EB相当于4462个美国国会图书馆的所有馆藏数据量。

大数据带来的经营理念的转变

一部关于奥克兰运动家棒球队的电影《点球成金》中就有球队通过统计学和数学建模的方式来获得比赛胜利的桥段。其中有一个非常有意思的场景。几个灰白头发的老球探坐在一起开始对球员进行评头论足。观众看到这里难免会有畏缩，因为人类在做决定的时候，往往非常草率。

其中一个球探说道：“这个很有天赋，不错，而且外在也不错。”

“他击打动作也很好，只要碰到球就会弹出去很远。”一个满头白发，还戴着助听器的老人附和着说。

还有另一个球探也说：“击打声音很大。”

“不过他女朋友不够漂亮。”一位球探打断了他们的谈话。

会议负责人听了后说：“这是什么意思呢？”

这个人听完肯定地说道：“女朋友不漂亮就说明缺乏自信啊！”

负责人似乎很满意这个答案，说道：“很好！”随后会议继续。

玩笑了一会儿之后，之前始终沉默的一个球探说：“这人的气场很强大。我要说的意思是我在他上场之前就已经感受到了他的气场了。”

“他的长相也不错，通过了长相测试。他需要的只是时间，因为他随时都能成为一名好的球员。”还有一个人附和着说道。

那个经常都有不同意见的人一直在说：“我的意思是他女朋友的长相实在一般。”

这场景中的所有判断都说明了人类判断的误区，事实上，这个推断

完全是在毫无实际标准的情况下做出的，尽管它披上了理智讨论的外衣。一个拥有几百万美元身价的球员，决定是否签约似乎也没有客观的依据，而是依靠感觉。这个场景虽然只是电影中的场景，不过日常生活中也非常常见。之所以说此场景有趣，正是由于其具体现实中普遍存在的特点，无论是曼哈顿的会议室，美国总统办公室，或是街角咖啡馆，各种凭感觉推断的情形皆存在。

电影《点球成金》中的故事来源于迈克尔·刘易斯的小说《魔球——逆境中制胜的智慧》。小说讲述的是一个关于奥克兰运动家棒球队（还有人称之为绿帽队或是白象队）真实的故事，主要介绍了棒球队的经理比利·比恩（Billy Beane）的经营哲学，他在经营球队的过程中摒弃了以往挑选球员的传统方法，换用了一种依靠数学建模和电脑程序的数据来挑选球员的做法。他的这个方法看似很奇怪，有些类似于“上垒率”的标准，而不是以往所用的“棒球击球率”的惯用标准。这方法虽然挖掘了这个项目的一个方面，却忽略了另一个始终存在的方面。球员的上垒其实并非最重要的，不论是地滚球还是三垒跑，无非只要是上垒即可。因此，当比赛非常精彩，可是数据表明透垒不是太管用的时候，比利也不会太关心这项中看不中用的技能的。

比利所推行的“赛伯计量学”（Sabermetrics）尽管招来了众多的质疑，但还是在奥克兰运动家棒球队中被严格执行着，它的命名是来自在美国高级棒球研究协会（Society for Advanced Baseball Research）工作时的体育新闻记者比尔·詹姆斯（Bill James）。一直到现在，在美国高

级棒球研究协会当中始终存在着非常奇妙的亚文化中心。比利和当年用“太阳中心说”来挑战天主教权威的伽利略一样，打破了所有惯例。最后，比利的奥克兰运动家棒球队在2002年的美国联盟西部赛中摘得桂冠，此外还取得了20场不败的骄人战绩。也就是从那时起，球探不再吃香，取而代之的是统计学家，他们成为了成了棒球专家，不少球队都将“赛伯计量学”用于球队的运作。

大数据所做出的贡献还包括了从依赖自身判断转化为依赖数据做判断。统计学家和数据分析家的出现使得很多行业专家和技术专家的光芒黯淡了许多，前者摆脱了传统观念的束缚，依靠数据进行统计分析得出判断。在数据相互关系基础上，判断是不会受到偏见和成见的影响的，这和莫里中校从来不认为在酒吧中喝酒的干瘦船长所说的航道信息为真是一样的道理。他们的判断来自收集的所有数据背后隐藏着的信息，也就是说一切分析都有根有据。而莫里的方法当中对于风向和水流为什么如此并没有具体解释，可是一次安全的航海旅程，似乎“去哪里”和“是什么”远比“为什么”重要得多。

现在，各种专家在自己领域的影响力逐渐削弱，这是我们每个人都能看得到的。媒体界，如“赫芬顿邮报”（Huffington Post）和高客网（Gawker）网站上的新闻已经同编辑的新闻敏感度不再有关，而是来自数据的结果。相比资深的记者，数据选出来的新闻似乎更合大众口味。一家名为Coursera的网络教育公司，对其收集来的数据进行深度分析，如哪个讲座的片段是学生重放最多的，他们会找出其中的原因，或讲得

不够明确或是不够吸引人，再将最终的结果反馈给课程设计团队。这在以前是难以想象的，而现在教学方法也因此必须发生改变。如前文提到的，贝索斯在发现销量增加由算法来推动的时候，就已经放弃了公司的书籍评论员了。

这说明要在职业领域取得成功，必须做到与时俱进，只有做到这一点的员工才是公司最想要的员工。安大略的麦格雷戈医生所采用的治疗方法来自近十年以来病患数据进行电脑分析处理之后所推荐的，尽管她不是医院医术最为高明的医生，更不是产前护理的世界权威，但总能通过该方法为早产儿提供最好的护理和治疗。事实上，她是一名计算机科学专业毕业的博士。

事实上，最初在大数据上尝到甜头的人似乎都不来自他们做出巨大贡献的领域，他们可能是数据分析家、数学家、统计学家、人工智能学家等等，可是这项技能却在他们的推广之下运用到了各个领域当中。Kaggle的首席执行官安东尼·戈德布鲁姆（Anthony Goldbloom）曾提到，通常来说，在大数据竞争平台上获得成功的人和自己做出成绩的领域总是不相关的。

曾有英国的物理学家差一点就通过设计一个算法系统用于预测保险索赔和二手车质量问题之间的关系而获得成功，还有一个来自新加坡的精算师，也通过预测人体对化合物的生理反应项目大获成功，在谷歌的机器翻译团队中，很多工程师翻译出来的语言甚至他们自己也不懂。除此以外，还有来自微软翻译部门的统计学家，他们经常谈论的居然是只

要有语言学家离开他们的团队，他们所得到的翻译质量就会提升一点。

因此行业专家的主导地位会逐渐被动摇，只不过他们不会完全消亡。未来同他们平起平坐的是那些大数据人才，这就好比是清高的因果关系和卑微的相互关系两者终于站在了同一高度一般。关于知识的价值要重新审视，尽管从前专业人才似乎要比全才更受重视，大众普遍认为财富来源于深度。可是这或许是小数据时代的产物，专业技能就代表着精确性，那时候人们还无法掌握和获得足够多的数据，也不够准确，所以直觉和经验都是判断的依据。那是个经验先决的时代，唯有从书本和别人口中，以及在潜意识中的知识积累才是做出明智决定的最终依据。

当海量数据开始冲击人们的时候，人们就可以从数据当中挖掘出更多的信息。大多数数据学家都将以往的方式视为迷信，这并非他们不屑，而是他们手中掌握了从前不具备的财富来源——大数据。作为一个外行人，行业内本身存在的争论无法将他们束缚，他们同行业专家不同的是他们绝不会因为自己支持的那一方观点而导致偏见。这样的结果也证明了衡量员工对公司所做的贡献多寡的标准也在发生着改变。这也意味着每个在职场上的人，要学什么，了解什么，甚至是为职场要做什么准备等等都有了改变。

现代工厂创建的基石是数学和统计学知识，有时候甚至仅有少许的编程和网络科学知识，这和百年之前的计算能力或是更早以前的文学很像。各种和思维类似的同行交际不再是人类价值的体现，更多的则体现在了同各类不同人的交际层面上，只有这样知识的传播才会更为深刻和

广泛。从前成为优秀的生物学家先要结识众多生物学家的标准没有改变，只不过现在除了要具备深度的专业技能外，还要有大数据的广度。一个生物难题要解决，可能只要和天体物理学家或是数据试图设计师联系就不成问题了。

大数据的人才在电子游戏领域似乎最早同行业专家平起平坐，他们联手要改变这个行业。这个行业每年的净收入已经超过了100亿美元，这甚至已经超过了好莱坞的票房收入。过去游戏公司设计出一款游戏，就指望通过它来赚取高额利润。当然公司主要考虑的还是它的销售情况，或者以此推出升级版或是在此基础上推出新游戏。设计师个人的能力催生了游戏的人物、情节、物品和事件的设定，他们仿佛是在用米开朗琪罗画西斯廷教堂的画一样认真地对待自己的工作。只不过这项工作绝非艺术，而是一门科学，所以艺术家所倚靠的直觉和情感用在这个工作上是不合适的，就如同《点球成金》里的那些球探一样，他们真的已经过时了。

类似zynga推出的FarmViile，FrontierViile，FishVille以及其他一部分游戏都是交互式的。Zynga表面上是从这些游戏当中获取各类用户数据，还基于这些数据修改游戏，事实上游戏的版本非常多。公司从游戏当中收集各种数据，只要有玩家通关出现困难或是某一个关卡不对劲不愿意继续的话，数据就会为zynga发现这些问题提供依据，他们随即可以进行修改。隐藏得更深的是zynga还会针对不同用户的特点设计不同版本的游戏，类似FarmVille的版本就有几百个。

公司的分析家们观察玩家是否发现有自己的朋友使用这些产品的方式是通过颜色，他们还可以了解虚拟产品的销量是否真正增加了。例如FishVille的玩家购买透明鱼的数量已经远超过其他产品6倍的时候，公司就会利用大量出售透明鱼以获得更多的利润。在Mafia Wars中，根据数据统计显示，玩家更为青睐带金边的武器和纯白的宠物老虎。这些信息都来自数据的统计，而非一个设计师在工作室当中所能发现的。zynga的首席分析师肯·鲁丁曾经提到："表面上看我们是一家游戏公司，但实际上我们做的却是分析公司的事情，因为公司所有的运作都基于数据而进行。"

显然这是一个具有非凡意义的转变，从前人们习惯用经验、回忆和猜测来做决定，正如W.H.奥登（Wystan Hugh Auden）在自己的诗中提到的："知识已经退化成了混乱的主观臆想，那营养不足是来自太阳神经丛的感情。"马萨诸塞州的巴布森学院商科教授托马斯·达文波特（Thomas Davenport）曾写过多部数据分析著作，他将其称为"黄金般的直觉"。执行官们做决定的时候所依赖的是自己的直觉。可是随着预测性分析和大数据分析越来越影响管理决策后，直觉所起的决定性作用就会彻底转变。公司无论是考虑出品电影还是签下哪个艺人，整个过程都会有本质的改变。麻省理工学院商学院的教授埃里克·布伦乔尔森（Erik Brynjolfsson）曾与自己的同事们进行了一项专题研究，研究结果表明依赖数据进行决策的公司比传统依赖直觉进行决策的公司运营情况要好得多，生产率可以提高将近6%左右。从竞争力来看，随着越来越多的公

司开始引进大数据，依赖大数据的公司竞争力会显著增强。

大数据的舆情服务

社交网络中人们的言行不但使尼尔森等市场调查公司感兴趣，政府机构也同样对此有强烈的兴趣。古往今来，治理国家很重要的一点就是要关注社会舆论，这是政府了解治理效果的一个重要渠道，更是发现人心向背的关键。所以不论是什么样的政府都对公共舆论十分关心。这一点也催生了另一种新型服务——“舆情”，即分析公共话题的热点、趋势以及对此的对策等等。

现代社会，人们闲聊、传话的重要媒介就是微博。从前人们的习惯是在街头巷尾与人闲聊，如今的人们通过微博来发表自己的行为，有时候打个喷嚏全世界的人都会知道。例如“微博女王”姚晨，她的微博粉丝有3000多万，有时候简单的一句“早上好”转发量就会高达2066次，还会有3739条的评论。这不过是即时的一个数据，在书籍出版之后，这个数字还会不断增加。

微博成为舆论场源自其巨大的传播和扩散效应。中国人民大学舆论研究所的一项研究表明，2011年舆情最大的信息来源就是微博，已经达到了20%以上。从此各级政府就以微博作为自己关注舆论走向的主要阵地。

同样应运而生的还有舆情服务。目前很多公司在提供舆情服务，譬如上市公司中的拓而思，人民网等等。通常来说省政府购买了舆情服务后，市政府也会跟着买，这是由于市长想在省长之前获得舆情信息。同

理，县政府，包括大一点的镇政府都会紧接着来购买舆情服务。一般公司提供的舆情服务都相对简单，通常是定期为政府提供网上的舆论热点分析，很简单的一份材料而已。

如果只是基本的报告，显然无法在舆情服务产业的升级过程中赢得胜利。在复杂严峻的舆情形势面前，企业和政府所需的还有舆情监测、舆情预警、舆情分析报告、应对处置、顾问咨询、舆情培训等多重服务。最高级的舆情服务应当是从现有的大众言论中挖掘接下来可能出现的舆论热点，提前介入其中，做到以恰当的方式来引导舆论，这才是防患于未然。“事后诸葛亮”无论如何都比不上干脆让某件舆情事情不发生，这就是上医治未病的道理。

舆情服务的最高级别必须依靠大数据分析等相关技术，少了大数据的支持一切就都实现不了。中小舆情公司之所以难以升级成功，最大的技术难度就在于此。高级舆情报告在于即时性、全面性和前瞻性。缺少数据中心、缺少采集、分析大数据的技术以及成熟的舆情分析团队的话，要达成如此高质量的舆情反应几乎是不可能的。

舆情若是从“信息聚合”角度作为一个产业的话，当中定有无限巨大的空间。消费者对产品、公司、品牌的意见和反馈都包括在舆论当中，即便出现了负面的舆论也会成为公司改变公众形象的重要机会。所以说公司潜在客户或者潜在广告受众，他们的舆论可以是正面的，也可以是负面的。就此而言，譬如尼尔森等市场调查公司、拓而思这样的舆情服务公司，未来的发展空间是巨大的。实际上，舆情服务不仅限于技术

活，更是一个跨多个学科的综合服务工种。它依赖大数据统计和分析，同时如果想在这个领域有所作为的话，还需要社会学家、心理学家、传播学家、数据科学家的共同服务。

大数据预测你的下一步行动

法庭上总是要求个人对自身行为负责。审判员在经过审理之后才会做出最公正公平的判决。可是到了大数据时代，为了维护个人动因想法，公正的概念要重新定义了——人们自由选择自我行为。简而言之，个人要对自己的行为而非倾向负责，这是应当做到的。

这自由权利在大数据之前是显而易见的，而这确实明确到不需说明。事实上法律体系当中总是通过人们过去的行为来判断其是否该为此行为负责，这是其运作的规则。只不过在大数据时代人们的行为就可以预测了，并且预测的结果是非常准确的。这使得我们对人们的评定不再依赖实际行为而是预测的行为。

确保了个人动因，政府对个人行为判定的基础是来自真实行为而非只依赖大数据的分析这点我们就可以确认了。因此政府追求的不能是依赖大数据分析而预测到的未来行为，而只能是过去的真实行为。再或者，政府对过去行为的追究过程中也不能只单纯地借助于大数据分析。比如借助大数据分析可以对两家涉嫌价格操纵的公司进行大致判定，随后监管机构再使用传统的方式对此立案侦查。显然，大数据的作用在于可以预测其是否可能犯罪，而不能对其是否有罪进行判定。

政府领域之外也适用相似的原理，例如关乎个人利益的公司重大决

策——雇用和解雇，再有就是按揭和信用卡的是否准发。假如要单纯依赖大数据做出决策的话，前提必须是有非常到位的特定防护措施。

第一原则是公开原则。由于结果会直接影响个人，所以用来预测分析的数据和算法系统必须公开。

第二原则是公正原则。算法系统必须是经由第三方专家公证过的可靠、有效的系统。

第三原则是可反驳原则。个人对预测进行反驳的具体方式也要明确（这一点和科学研究当中必须披露所有可能影响最终结果的传统很相似）。

确保个人动因防范“数据独裁”带来的危害是最为重要的，而这危害正来源于数据被我们赋予了原本没有的意义和价值。

同样重要的还有保护个人责任。正因为这一点的吸引人，无论社会做出哪一种和他人有关的决策，决策者们都不需要再承担责任和风险。反之，风险管理成了管理的重点，实际上就是对可能性所进行的风险评估。所有看起来客观的数据，可以帮助去除决策中的情绪化和特殊化的部分，评价者主管的评价被数据运算法则代替，而决策的严肃性也无须通过追究责任的方式来显示，更准确地应该称作“客观”的风险和风险规避，这主意怎么听都感觉不错。

例如，那些被预测可能犯罪的人因为数据分析的结果被隔离，不断审查他们只为规避风险，可是事实上他们所受的惩罚来自并不需要他们承担的责任。设想一下，一个青少年被“预测警务”的运算法则预算出将

来五年他可能会犯下重罪，于是政府部门就开始对其例行监视，每月都会有一名社会工作者去拜访他一次，目的是为其解决问题。假设少年身边的亲属、朋友、老师和雇主认为这种做法很是耻辱（这种情况发生的可能性很大），那么很显然这拜访就成了一种惩罚，是对尚未发生的事情的惩罚。不过要是不将其视为惩罚，而是认定为一种对未来风险的规避的话，也就是说把风险降至最低点的做法（这里所说的最小风险是破坏公共安全的最小风险），也未必见得就是一件好事。社会用干涉、降低风险的方式来代替自我行为负责的方式会导致个人责任意识的贬值。保姆式的国家才会主张预测。实际上，对个人行为所承担责任的否定就是对个人自由选择行为权力的摧毁。

国家所做出的决策如果均来源于企图通过预测来规避风险的愿望，那么个人的选择就不存在了，自主行为的权利就更不用提了。无罪，无清白，这样一来不会带来世界的进步，而是倒退。

数据也会骗人，从人的动作推导数据

美国电影《致命魔术》在片头就向观众提出了这样一个问题：“你真的看见了吗？”不少人认为自己眼见为实，可惜并非真正的观察，事实上他们没有看到真相。

“观察”究竟是什么？从事物内部寻找其相对性就是观察，简单说就是在共通性中寻找相对性。因此，一个人所说的话是真是假，如果要判断一定要基于以下两种情况：一是看，看看这个和自己说话的人是不是第一次见；二是看看日常生活中这个人是不是已经见过多次面。这些对

于判断一个人说真话还是说假话有着独特的意义。熟悉的人要判断是否说谎，和判断一个素昧平生的人是否说谎相比要简单很多。问题在于后者无经验可借鉴，短时间的观察确实很难从中提取相对性。

说到这里，有两个词语不得不先了解一下：常态和时态。先理解“情态”再来说常态吧。所谓情态，藏在体内是谓“情”，展现在外的是谓“态”。《礼记·礼运篇》就曾提到人有七情，即喜、怒、哀、惧、爱、恶、欲，这些都是人内心的情态。体内有了这七种情态交织在一起，当人们的边缘系统被激发了之后，时态就会衍生出来。那些因为赌赢而欣喜若狂的人，正是边缘系统出了问题。一旦冷静下来他就会后悔刚才的表现：“太失态了，太丢人了。”

一个人的时态是最容易在观察一个人的时候发现的。一个人时态的差异，要从了解这个人的常态开始。什么是常态呢？通常有以下六种：弱、狂、哗、周旋、慵懒、媚。弱态是指那些动作温柔、说话轻声细语，有很强包容性，如小鸟依人一样的常态。狂态则是坚强好胜，不修边幅，言谈举止都如若无人在场。身边朋友一对比就会发现他有什么样的常态。问一个具有狂态的人的意见时，他总认为自己是对的，也愿意把自己的观点表达出来。弱态的人则不会这么做，意见是模棱两可的。哗态的人则大多在遇事时大手一挥，说道：“我来说，你们都别说。”周旋态的人在选择面前总有太多的纠结。慵懒态的人表现常常是非常漫不经心，不在乎一切。媚态则常常出现谄媚的行为。

日常生活中，人们会表现出不止一种单纯的态，会有多种态混合在

一起。可是不管如何，凡弱且媚的人一定要远离，凡狂且媚的人，通常都比较了不起，这是一条交友法则。

一个人的常态是日常判断他的基本依据。如果要向一个慵懒态是常态的人汇报工作，发现这人突然一反常态地正襟危坐的话，那必然是非常重视这件事情。

判断是不是谎言，常态和时态必须区分。一个人的常态该如何了解呢？观察是最有效的方式。上文提到过判定熟悉的人是否说谎难度比较小，正因为了解对方的常态，而不了解常态的陌生人要做相应的判断难度要大很多。

日常生活当中还会听到变态这个词，事实上，这就是常态向时态转化而出现的异常举动或是不理智的行为。要是对常态和时态都非常熟悉的话，那么要观察一个人变态的起点和变化点就不算太难。

除此以外另外一点也是观察的重点。一个人的常态在变化的时候，是什么环境或是事件促使他变化的，这是必须思考的。譬如一个被警察抓到警察局的人，他必然会焦躁不安，且十分恐惧，动作也会变得粗鲁，但这不说明他必然是恶人，因为在判断时还要考虑环境因素。若是发生了一起爆炸案，那么究竟有没有什么异常的人在爆炸发生前后出现在现场呢？案发前是不是有异常逗留在现场的人呢？案发之后，如果有人受伤，很多人都凑过来救助，当中有两个行色匆匆，且表情淡定的人经过，那这两个异常的人就有很大的嫌疑了，无疑他们就是常态中的变态。

从人推“数”，观察数据时有以下几个问题不容忽视：常态的数据和时态的数据是什么？什么是骗人的数据？直观感受和细心观察后常常会得到大相径庭的结论。

依照通常的理解，一个人在注册某一网站填写性别的时候，不是男性就是女性。可是在阿里巴巴，谁曾想到性别标签居然有18个。

经过仔细的调查，用户性别的识别方法就浮出水面了。例如一般习惯在早上浏览的就偏男性一些，晚上浏览的则偏女性一些（很可能是妻子在使用）。真实中的性别仅有1和0之间的关系，现实却存在0-1的关系，70%的是男性，而30%的是女性。不同的场景性别会因为不同的原因而发生变化，例如搜索、社交和广告都会带来变化。“真实”性别在静态时，通过A/B测试的表现和动态虚拟性别相比有效性更差，这也正好说明了数据必须是运动变化的，绝不能仅是闭门造车，阿里巴巴类似这样的案例实在太多了。

说到数据化运营，通常在公司内部流程是这样的，比如一个童装项目需要通过营销推广来招揽新客户。第一步要做的就是定位客户群，努力找出所有对此项目感兴趣的消费者标签，向他们发送邮件或是短信引起他们的注意。而这一流程便是数据化运营，企业利用这方式来解决问题。最为简单的解决方式就是找出所有曾经对此感兴趣却没有产生消费行为的客户群，再比如用关联类目来扩大自己的目标客户群。

“用”升级为“养”，这就是运营数据，阿里巴巴用这种方式在淘宝用户中寻找。举12岁以下儿童商品为例，淘宝会在自己所有的用户中搜索

家中有12岁以下孩子的消费者，他们的重要特征是从未有过此类目商品的购物行为。这样一来，从前只在有过购买行为的用户中运营，如今一下子用户就会扩展几十倍，从几百万到几千万，运营数据的结果就是这样，它从主动收集手机数据开始，到运营数据，最后的结果是产生新的数据。

大家广泛理解的数据运营本身就是个不断的循环。循环的过程当中，总会产生不少新的、不同维度的数据，而在这个循环当中，数据也在不断地适应，这才能运用到数据化的运营当中，将从前的运营方式彻底改变。大数据落地的方式或许正是如此。

假定数据是稳定的，这是数据化运营的前提，这也可以用来改变企业的运营模式。从前唱主角的是结构化的数据，例如绝大多数的企业总是在企业运营中使用财务数据，多年沉积下来的财务数据相对而言比较稳定。可是在数据化运营下，由于非结构性数据和半结构性数据的产生，难度就增大了，例如音频就是典型的非结构性数据，而且它也很难转化为结构化数据，这样的数据产生之后，企业运营若是依赖数据的话，其中的技术难度可以想象。

稳定的数据是进行数据化运营的基本前提，相比之下，运营数据则需要可获取的数据，它们必须是不稳定的。

这话该怎么解释呢？比如说关键词ROI，尽管这是一个听起来很稳定的数据，但是其反映的却是短线利益。商业是数据化运营要紧紧依存的对象，数据和商业混合在一起，再用假设稳定的方式进行业务对比和

细分，还有趋势估计。

这不代表稳定就全是对的，毕竟在企业对数据的运营分析中很多稳定因素都需要假定存在。

网络数据背后的价值

很多具体领域中，企业要了解它们的顾客的途径是网络数据。驾驭不来大数据的企业是无法洞察他们顾客的特点的。接下来就来好好明确一下能通过网络数据来洞察的领域。

先来说说购物行为。顾客为何会进入这个网站并开始购物行为，这是理解顾客购物行为的最佳起点。顾客使用的是哪种搜索引擎，搜索时输入的是什么关键词呢？他们之前收藏的书签又有哪些呢？这全部的信息是分析专家寻求模式的依据，譬如产品的销售量同哪些关键词，哪些搜索引擎，还有哪些推荐网站彼此关联。分析专家更为重视的不但是要查看哪些产品在给定的网页中拥有高销售额，更要看看哪个时期同一顾客的购买行为更多。网站的销售和顾客的购买行为两者跨渠道地结合就是最终的价值。

顾客登录网站先要做的就是浏览所有看得到的商品。专家在处理数据的时候要做的是甄别顾客是在看了哪些商品的登录界面后离开的，而还有哪些顾客会继续浏览，哪些人又继续看了附加图片，阅读了产品评价，浏览了产品的详细说明，看了运输信息，又利用了网站上的其他可用信息等等也要了解。就比如，要鉴别一下进行比较的是哪些产品，最后鉴别出来的商品又是如何被移入购物车，或者是被从购物车中移出。

网络数据是无可替代的，顾客接下来会购买什么，又是如何进行决策的，它都可以告诉人们。了解了这些显然可以帮助企业促使顾客完成尚未完成的交易。顾客由此几乎会认定自己在购买商品的时候，企业已经了解了他们的想法。

网络数据还有一个非常有趣的功能，顾客在购买行为之前会因为它发现自己感兴趣的商品被捆绑在一起。在购买行为完成以后再为其推销商品的做法已经过时了。而现在的做法是在第一时间就让顾客浏览到自己所要查看的商品，还将其感兴趣的商品捆绑起来。

就例如有一个正在浏览电脑、备份硬盘、打印机和显示器的顾客，看起来他是想要升级自己的PC系统。很显然，商家就要为其提供一整套顾客浏览过的部件套装。不能在顾客已经决定了购买电脑后再提供这些信息。购买之前，为顾客提供的定制捆绑优惠策略远胜于那些在购买行为之后推荐通用配件的做法。

再来说说顾客的购买习惯和途径。网络数据可以收集、分析和查看顾客在网站上的浏览历史，也就能把顾客购买决定之前的过程都弄清楚，从中还能发现顾客的购买偏好。以航空公司为例子吧。顾客预订的机票可以帮助航空公司来了解顾客的偏好，其中包括顾客会提前多久预订机票，预订的是哪些舱位，是不是跨过周末的航班。所有的信息对于航空公司都是十分有用的，网络数据能让他们获得更多的信息。

航空公司从中能发现看重便利性的顾客是哪些，通常这些顾客的做法是找到特定时间的直飞航班。除非是价格悬殊，但便利性的获得却只

有一点点的时候，他们才可能放弃便捷的直飞航班。一个在纽约JFK机场转机的乘客，如果选择直飞LaGuardia机场会节约700美元，而这期间他需要耗费30分钟在JFK机场，额外的出租车费用仅仅是20美元。如此情形下，重视便利性的顾客会选择在JFK机场转机承受一切麻烦，只为了省下700美元。假使差价仅仅是200美元，且放弃的话到达目的地的时间要晚2个小时左右，那么同样的顾客他所做的选择就会是最便捷的直飞。

哪些乘客价格至上航空公司也可以分辨出来，他们的习惯是从中搜索出价格最优的航班。只有一种情况会让他们违背价格优先考虑的原则，那便是可以用牺牲适度的价格来提高便利性。譬如，一个顾客选择上午10点出发，航班的价格是220美元，6点的航班价格是200美元，如果要让他牺牲四个小时的睡眠来省下20美元的话，显然他还是会放弃价格至上的原则，愿意付出20美元的溢价。

航空公司基于搜索模式就可以了解顾客的购买习惯是偏向价格还是偏向特定的目的地。一切在于顾客是否对所有特价机票都有所了解的情况下才选择了其中的一张，还是只是看到了某一特定目的地的机票就立刻购买呢？一个大学生，对他来说春季假期旅行的目的地，很多地方的意义都是相同的，那他一定会先选择最为优惠的那个。可是对于定期回家的顾客来说，目的地才是他最为感兴趣的因素。

企业如果能了解顾客最经常在网站上浏览哪些目的地的周末特价机票的话，那要推测出顾客的购买偏好并非难事。很多顾客随时都有回家

的时间，那么他们就会很是关注飞往特定城市的特价机票。只要有特价机票的话他们就会立刻预订。航空公司通过识别这个模式，就可以很好地预先了解顾客的需求。

所有上面提到的这些例子，其实都是在考虑怎么把现在浏览和研究的模式，与以往的历史经验和购买记录彼此结合，带来更多的惊喜价值。只不过如此模式的分析流程要改变的话需要耗费大量的时间和努力。只是网站吸引不同顾客的方方面面被知晓后，就可以很容易地向已经被定位的顾客发送他们最感兴趣的信息，充分地满足顾客的需求。

再说说研究行为。要和顾客个体进行交互的话，前提是要理解他们是如何利用网站内容的，除此外还可以帮助理解网站的哪个部分因此获得了销售的提升。顾客在购买过程中所探究过的选项是推测他们重视的因素的重要依据。

就举一个专门做电影销售的在线商店为例。假如顾客在做出购买决定之前首先习惯去查看视频有哪些格式，像是有标准格式、宽屏、扩展或是高清格式，也就是说绝大多数时候，他们更青睐某一种格式的视频。但事实上，他们对格式并不在乎。那么，网站可以每一次都为查看过所有格式视频的顾客提供不同的版本。那究竟是为什么还要让顾客在众多的版本中挑选呢？事实上某一类的顾客他们已经习惯了购买一种特定格式的视频了。

要对顾客进行研究的话，利用网络数据还有另一种方法，那就是在网站上的所有信息中鉴别哪些对顾客，特别是重要的顾客有重要的价

值。在购买之前，顾客会有多高的频率去查看评论、附加图片和技术说明呢？只要跟踪不同的会话，再结合顾客的其他数据，就能了解顾客购买的整个流程，在哪一天进行查看浏览，哪一天完成购买行为。最终的购买行为已经是指向十分明确的网络会话了，因为交易因此而完成。网页浏览的历史可以拼出一张完整的图像。很多时候企业试图去除网站上的一些极少被用到的特性，可是这些对某一类重要的顾客来说确有十足的价值。如果是这样，这些特性就不会被剔除掉。

至于顾客在购买行为发生之前到底对自己想要购买的商品进行了多少调研，这些也不再依赖昂贵、小型的调查了。网络数据能够为企业分析出每一个顾客个体或是群体最重视的因素，并且还能规避掉很多顾客言行不一带来的风险。这样一来真相就显现出来了。

顾客的一些异常行为会被企业所发现，有时候顾客会详细查看某一产品的说明，随后又没有购买，反而是那些未曾看过产品说明的顾客却没有这样做。很可能是因为在查看了产品说明之后，从中发现有些描述还不够详尽，或者是说明中还存在疏忽等等，总之升级产品描述的话无疑对提高销售业绩是非常有帮助的。

其中有一个重要的指向标，那就是阅读评价，顾客所看重的特征都可以从中发现。顾客看重的是哪些评价，又不在乎哪些评价？哪些商品在顾客阅读了商品评价之后就不再购买，又有哪些商品评价是可以帮助提高销售额的？一旦发现有人在阅读了某些评价之后放弃购买或是决定购买某商品的话，那这些评价就要留心了。也许会有负面的评价出现，

而对于网站而言核实这些评价是必要的，他们提出了哪些问题，就要想办法去处理掉这些问题。

最终顾客所看重的网站特性也要弄清楚，还有顾客在网站上如何浏览商品也是必须了解的，网站只有这样才会贴近顾客。那些重视详细产品说明的顾客，必须让他们一看到商品就会立刻去浏览产品说明。而那些习惯去浏览图片的顾客，必须为他们提供全尺寸的图片而不是缩略图。最重要的是要为顾客提供搜索和查看的便利，只有这样顾客才会在做购买决定的时候选择自己而不是选择其他商家。

最后来说说反馈行为。对商品和服务的详细反馈信息是顾客为网站提供的信息中最有价值的。顾客确实很愿意为自己所钟情的品牌去做这件事情。从文本当中发掘顾客反馈时所用的语气、意图和主题，就能明白顾客之所以看重它的原因了。

顾客在购买商品之后，会不会照惯例去发表评论呢？评论如果是积极的，还有其他的很多顾客阅读到这些评论，这些顾客从中所获得的特殊激励显然是公司最明智的做法。其他类似的还有顾客在线求助会话中的问题和评论，通过分析就能知道顾客普遍要了解的是什么，也可以知道特定顾客需要了解的是什么。特定顾客所看重的特征一旦被发现，那就可以有针对性地为其推荐相类似的商品了。

通过顾客在各个渠道上发表的评论和问题，可以了解他们的好恶，还包括譬如他们是不是Facebook的粉丝，在twitter上他们是不是彼此关注？此外，网站如果发现了在各个社交媒体上积极地发表正面评论的顾

客存在，那么网站就可以着力将其培养成最具影响力的推广大使。企业品牌的影响力的提升，值得企业花时间精力去做这样的付出。顾客的影响力和个人的价值之间并不一定是密切相关的。比较活跃的客户一般都是享受标准待遇、规模中等的客户，实际他们为销售所带来的价值已经远远超过了他们的影响力，剩下要企业做的就是提升这些客户的待遇级别。

第六章 颠覆与重塑思维，大数据与思维革命

大数据所昭示的是一种颠覆性的思维方式，无论是人才的需求、支付方式，还是营销策略、发展预测等等，企业都要找到与之相匹配的经营思维方式。这注定是一场惊天动地的变革，而引导这场革命的正是大数据。

大数据时代的综合人才

人才是大数据时代最为稀缺的资源。谷歌的首席经济学家Hal Varian提到过，统计学家是未来最为性感的职业。大数据创新和创业最大的挑战就来自人才的招募、培养和使用。合理的模式将人才的价值放大，同时还将大数据的价值释放出来。在大数据时代人才缺乏的原因是复杂的，破解的方式也是多种多样的，为此大数据实验室做了很多有益的尝试。

不少企业的数据基础是很不错的，还有很多待解决的有意义数学问题，并且大数据战略被引入的强烈愿望也因此体现。限于企业规模或是企业发展的现状，总不会有那么多的企业可以和数据科学家合作，大数据战略便因此而搁浅。北京中关村，中国智力资源最为密集的地方，这里就有大数据实验室，在高校密集的上海杨浦区还有中关村的分部，有了这个，中国顶尖的数据科学家就可以和大数据实验室合作。不少在大数据方面有着大抱负的企业或是创业者他们看到了这种合作中，自己可以求助数据科学家为自己的大数据难题攻坚的机会。同时大数据实验室微信的多样数据科学家网络也将和大数据实验室共享，在人才成本高，技术需求多样的今天，这些问题也得以解决了。数据科学家其实对此也非常感兴趣，毕竟这些问题具有多样性的特点，且现实意义很深远。

大数据的创新和创业除了需要数据科学家外，还要有拥有专业知识和技能的人才来协助。其中有IT能力、行业知识、创业知识、投资知识等等方面的人才。在大数据实验室的支持下，具备多种专业技能的导师

（mentor）网络就可以得以建立和维系。大数据的创新需要来自各个行业的专家定期或是按需进行指导，大数据创新者的成功率也会因此提升。

此外，大数据人才从大数据实验室中获得了多种可能，他们可以由此来依照自己的特点来施展自己的才华。大数据人才一定是对大数据研究感兴趣的，和大数据实验室开展合作并以此来研究分析大数据是最佳的选择。如果大数据人才能兼备商业和技术才能，就可以通过大数据实验室的资助开始创业。

大数据人才展开了与大数据实验室的灵活合作，大数据人才会因此成长。数据价值实现就应当要依靠人才价值的实现。那么，我们应该思考这样的问题。

（1）各个国家的报业集团纷纷破产、停刊说明了什么？谷歌到底是互联网公司，手机制造商还是电信运营商呢？无法定位的谷歌，事实上它的利润来源还是基于大数据的精准广告。谷歌的数据量越大，广告就越是精准，利润就越多。谷歌是一个精美绝伦商业模式的开创者，用户可以免费使用它所提供的优秀服务，而谷歌做的只是用大家使用的数据向广告主收费。在大数据成功飞速运转的背后，就是谷歌的胜利。激烈的竞争压力下，传统报纸和杂志举步维艰也就不足为奇了。

（2）美国总统似乎是最善于营销的，他的营销在于数据驱动的营销，这种模式为美国历史上的第44任总统奥巴马的续任所起的作用是巨大的，同时也成了2012年大选中的一个决定性因素。这也传递出了一个

信号，华盛顿那些原本从直觉和经验中决策的竞选人士失去了原本的优势，反倒是依赖数据分析专家和电脑程序员的人们从大数据中获得了关键的优势，大数据提供了众多信息，从而使他们准确理解选举的形势。显然大数据在政治营销领域也迎来了自己的春天。

（3）庞大的传媒业市场，正在高速发展。数据资产综合应用了新技术、新商业模式、更多维度以催生下一个巨人。

飞利浦的大数据营销策略

世界上最大的电子公司之一飞利浦，它拥有多个领域的电子产品。飞利浦将中国市场作为“极有潜力的本土市场”，在整个营销战略中成了一个重点地区。所以，在大中华区，生产小家电的飞利浦优质生活事业部将其视为与欧洲、美洲同等级别的市场，成了“商务组织”的四大核心市场之一。

飞利浦最具优势的小家电有个人护理小家电、榨汁机、吸尘器和空气净化器等等。只是在中国市场上，还有美的等品牌对手带来的激烈竞争，飞利浦受到了极大的冲击。从2011年5月开始，在全国各个直辖市和重点省份，飞利浦小家电和精品小家电开展了一个季度的推广促销风暴，以推广自己的品牌和网络的曝光量。

执行项目期间，传播活动重点推广和促销了飞利浦空气净化器、风景时尚灯、吸尘器、剃须刀、soudbar、avent六大产品。扩大品牌和产品网络的曝光量是此次传播活动的主要目的，但同时还要让目标的消费群体更多地认识飞利浦的相关产品，这样一来商业效益就会随之增加。

中国互联网环境的特征是网络数据各方割据，有显著的碎片化情况。除此以外，飞利浦的产品线过于繁多，导致了传播任务十分繁重，飞利浦很难掌握单一品牌的互联网互动方式，这种方式也不利于其传播信息，品牌形象也难以树立。飞利浦的目标在于整合营销策略，中国消费者由此能更深地体会到飞利浦品牌的优秀特性，有限预算的ROI也会大大提升。

实际上，飞利浦投放广告要是从互联网传播角度来看必然会迎来四大挑战。第一个挑战是借由大数据的洞察，人群和精品人群的网络行为特征和心理特征也很快被洞察出来，网络传播就以此为策略依据。第二个挑战是媒体传播策略是在区域的销售策略基础上来制定的，线下销售和网络推广得以有效地衔接。第三个挑战是多产品广告同步推送的问题解决了，有限的媒体版位也得到了有效利用，广告版位通过技术手段来展示消费者最感兴趣的广告，不同的受众看到的广告是不同的，版位的价值因此提升。第四个挑战，在亿赞普ID数据的基础上，跨区域广告实现了频次调度，有限广告位产生了最大化的销售价值。即便是同样的预算却达到了2倍以上的常规广告投放方式效益。

传播策略的确定先要来自由大数据建立的数据模型，再依据飞利浦消费群体做出相关数据，可是在飞利浦的全国推广中如何运用大数据呢？

飞利浦项目的执行基础是海量数据的存储基础，再经过数据挖掘和人工智能的算法，分析海量的互联网用户、内容和相关行为，从中发掘

潜藏的营销机会，这才会收获最有价值和效率的营销效果，投资回报率才会更高。

执行策略的过程基础就是互联网的大数据分析，再同互动策略、数字创意、互联网媒体采购、互联网公共关系和监测服务全面进行结合整合，构建一系列关于差异人群覆盖、品牌植入传播、多媒体组合策略、EPR互动口碑传播以及CRM用户持续管理系统等完善的体系，在技术和媒体数据化结合的基础上提出基于智能化投放的360度传播策略，目标就是要全面覆盖飞利浦的全部受众。

这不单纯只是营销，还是数据、技术和营销的完美结合。

阿里小贷的“不可能的任务”

阿里巴巴在2012年8月对B2B小额信贷的业务（以下简称为小贷）范围进行扩大。以往只能是B2B平台付费用户的小贷业务，普通注册用户从那以后也可以申请小贷，地域范围却仍然局限在江浙沪一带。此前就无地域限制的淘宝网和天猫商城贷款业务并没有因此发生变化，这也在某种程度上意味着阿里巴巴小贷业务会激增，也会进一步提升它所占据的市场份额。消息一出，社会各界的反响都很强烈，只不过大家的意见并不统一，有人甚至喊出了“颤抖吧，银行！”的口号。事实上阿里巴巴的这项业务会颠覆传统银行吗？谁也不知道，毕竟最后的结果是多种要素综合的结果。可是现在有一点可以确定，在金融界，阿里巴巴的小贷业务确实有了自己的一片立足之地，传统银行业也因此受到了一定的威胁。

始于2007年5月的阿里巴巴小贷业务，其实是阿里巴巴在和中国建设银行、中国工商银行合作的基础上推出的中小企业无抵押贷款，但后来因为种种原因，这次合作并没有维持太长时间。2010年和2011年，阿里巴巴和复星集团、银泰集团还有万向集团合作成立了浙江阿里巴巴小额贷款和重庆阿里巴巴小额贷款两家公司，主营业务是网络贷款。总体来说，阿里巴巴的小贷有两大类，一种是B2C平台，也就是基于淘宝网和天猫商城客户的贷款和信用贷款，另一种则是针对B2B平台，为阿里巴巴中国站和中国供应商会员提供阿里信用贷款，具体区分的话可以分为循环贷和固定贷两种类别。小贷的贷款额度不超过100万元，计息的方式是按日收费，一般淘宝网、天猫商城的信用贷款和阿里巴巴的循环贷的贷款利率都是0.06%/天，剩下的贷款项目为0.05%/天。这样很容易可以推算出前者的年利率大约为21.9%，后者的年利率大约为18.3%，贷款的利息均高于央行1年期贷款6%的基准利率，属于小额贷款行业内中等水平的利率，在国家规定的基准利率4倍以内的要求。

中国中小企业60%以上的GDP都来源于此，80%左右人口的就业问题也因此解决，中国社会经济发展也由此推动。不健全的中国信用体系，造成了非常艰难的信用记录和信用评估，国内金融界为了控制贷款的成本高额风险，长期以来主要以抵押、担保贷款为主，信用贷款还是比较少。中小企业在创业初期既缺抵押物，又无担保，传统金融机构很难为他们提供更充分的融资。阿里巴巴的小贷既不需要抵押，也不需要担保，还有非常高的申请效率，对于中小企业极其迫切的融资需求来说显

然是一剂良方。适应市场发展要求的企业必须有强大的生命力，阿里巴巴小贷曾经创造出突出的经营业绩，它所凭借的就是创新业务。阿里巴巴金融统计数据曾指出到2011年年底，接受阿里金融服务的小微企业已经多达9.68万家，占到了全中国4000多万家企业的0.2%，投放贷款也高达154亿元，坏账率不及1%，和传统银行抵押类贷款产品的坏账率相比低了不少。阿里金融更让人惊讶的是仅仅2012年上半年投放给小微企业的贷款就有170万笔，总金额有130亿元。每天平均完成了1万笔贷款，平均的贷款额度在7000元。

阿里巴巴金融的单日利息截止到2012年7月20日高达100万元，从这个趋势来看再有一年的时间，就会实现一年3.65亿元的利息收入。相对而言一般的小贷公司一年的利息收入只会有几千万元。国家有政策规定，银行借贷的数额不得超过其注册资本金的50%。阿里金融的两家小贷公司注册资金的总量在16亿元左右，照此计算，放贷资金就不能超过8亿元，那么日利息收入最多的时候可以超过120万元（日利息为0.05%），上限大约是144万元（日利息为0.06%）。阿里金融现在的日利息大约是100万元，从国内来看算是业内顶尖的了。阿里金融的扩展还局限在自己的生态体系内部，不过凭借着它7980万注册用户量、1030万个企业商铺和75.39万名付费会员量（截止到2012年3月的数据），阿里即便是不算上淘宝网和天猫商城的庞大用户群，也可以随着自己生态体系的逐渐庞大推动自己金融未来的发展，这种势头不容小视。

既没有担保又没有抵押，对哪一家银行来说都是个大难题。如果还

要实现每日1万笔贷款，单笔额度在7000元左右，单日利息的收入要做到100万元，很多传统银行更是想都不敢想。阿里金融又是怎么实现上述所有的一切的呢？阿里巴巴的回答一定会是他们经营了多年的主要业务，即从电子商业平台上累积了多年的海量数据帮了个大忙。

传统的金融机构为了规避信息不确定所带来的运作风险和经营损失，不得不用抵押和担保等手段，可是大数据时代到来，所有信息都会日渐透明，商业运作的方式也会随着商业环境的变化而改变。经过十几年的运营发展，阿里巴巴旗下的阿里巴巴（B2B）、淘宝、支付宝已经有了大量后台数据的累积，除了有用户的交易数据，还包括了资金流动、访问量、产品变化、投诉评价、用户注册等经营数据和身份信息，即时生成的这些数据都是自动产生的，阿里小贷从中获得了海量的数据支持，更重要的是这种方式成本很低。阿里巴巴打通了旗下的所有平台，实现无缝连接，网络数据模式由此引入了海量的数据，在在线视频调查和第三方验证等手段的配合之下，评估企业和个人的信用等级。与此同时，阿里金融还有一整套涵盖贷前、贷中、贷后的风险预警和管理体系，这就是控制贷款风险的方式。阿里金融依靠这个方式在金融机构中做到了传统银行所做不到的事情。

阿里巴巴到现在更是认识到海量数据给予他们的巨大财富来源。阿里巴巴集团在2012年7月宣布设立首席数据官，也推出了“数据共享平台”战略。2个月以后，阿里巴巴的董事局主席马云通过网商大会的平台宣布2013年起阿里巴巴开始转型，要重新整合平台、金融和数据三大业

务。马云的话当中不难透露出数据在阿里巴巴众多业务中的重要作用。显然，在土地、劳动力和资本之后，数据俨然是价值巨大的新资产，未来的企业竞争将围绕大数据收集、挖掘和运用展开。

第三方支付业务的另一种思路

2010年6月起，央行前后四次发放了大约200张类型各异的支付牌照，涵盖了互联网支付、银行卡收单、预付卡发行与受理、移动电话电话支付、固定电话支付、数字电视支付、货币汇兑七大类型的业务领域。由此可以看出，第三方的支付行业已经迎来了“后牌照时代”，行业的准入门槛降低了，但市场竞争却越发激烈。第三方支付企业以往的利润均来自交易佣金，大致是交易金额的0.5%到1%之间。由于互联网支付市场一时间涌入了各方资金力量，短时间内盈利空间被大大压缩了。

第三方支付企业为了摆脱瓶颈开始积极拓展新业务，他们先后推出了理财服务、行业解决方案、移动支付、跨境支付结算等业务，大数据为基础的创新业务是其中最具潜力的发展方向。互联网的支付行业走过了多年的发展，有了海量用户数据和交易数据的积累，这些数据已经成了包含巨大价值的“金矿”。第三方支付企业未来最重要的核心业务很可能就是基于数据挖掘和加工的商业创新应用，传统的支付结算的地位会降为“副产品”，之所以还留存就在于它能够源源不断地为企业提供市场数据。

国内的第三方支付企业实际上已经认识到了蕴含巨大价值的大数据，也开始挖掘“金矿”的工作。阿里巴巴集团应该是这行业中走得最快

最前的企业了，它用支付宝把淘宝、天猫和阿里巴巴B2B旗下的几个平台都打通，还在网络小贷服务上收获了巨大的利润。除了阿里巴巴外，其他第三方支付企业也开始了与阿里巴巴不同渠道的探索和开发，譬如快钱，它主要集中在企业应收应付账款融资服务的研发，快钱和合作银行之间采用将企业的应收账款或应付账款信息和产业链上下游企业一段时间内的资金流转数据统一的方法来合作，企业客户就可以以此从银行贷到款项。还有易宝支付也尝试着和银行合作，主要为航空领域的代理人提供周转资金，还在交叉营销业务进行新的尝试，通过这两个业务，易宝支付迄今为止的年收入大约是10亿元，在线支付的收入占80%，剩下的20%是营销和信贷业务。上述每一家公司的信贷创新都和第三方支付企业所积累的海量数据有关，显然经营发展的关键资产就是海量数据。从现在的趋势来看，未来基于大数据的新应用和新服务只会越来越多。

社会逐渐认可了第三方支付的独特业务模式，大量的买卖双方数据可以慢慢积累，也因此让第三方支付业务成了受众认可的合法业务。将来假如想继续推进其独立发展的话，下面两个关键点需要关注，一个是为了构建巨大的数据源必须通过创新产品和服务来吸引更多的客户，另一个就是扩展新型的业务，推动数据资产的开发和利用。

P2P网络借贷动了谁的奶酪

美国的Lending Club在新一轮的融资中，于2012年6月募集到了1750万美元，这当中包括了KPCB投资的1500万美元，还有摩根斯坦利前

CEO John Mack以个人名义投资的250万美元。有了这两笔投资之后，KPCB合伙人，被人称作“互联网女皇”的Mary Meeker和John Mac都成了Lendjng Club的董事会的董事，从那以后，Lendjng Club的融资已经接近1亿美元。这两位来自互联网领域和金融领域的大腕为何如此关注Lendjng Club呢？这必须从充分了解这个公司开始说起。2006年成立的Lendjng Club，于2007年正式上线，它作为一个网上金融社区，聚集了大量的优秀贷款人和借款人，作为双方借贷关系的中间服务方，它在尝试用一种相较于传统金融机构服务更为高效、便捷且成本更低的运作方式开展业务。

Lendjng Club平台上所设定的利率是不断调整的，依据贷款等级和期限的不同在6.03%到24.89%之间上下浮动。Lendjng Club平台上的贷款有A到G七个等级，一个等级下还有1-5五个等级，贷款时间上限为36个月。Lendjng Club向借贷双方收取中间费用，这是它利润的主要来源，每一笔交易成功后，借款人必须支付贷款总额的1.11%到5%不等比例的费用给Lendjng Club，具体金额由贷款等级和期限来决定，Lendjng Club随后再向贷款人收取借款人借贷总额的1%的服务费。倘若借款人的借贷服务不成功，Lendjng Club还是要收取15美元的费用。倘若借款人偿还贷款的时间超过了期限15天以上，Lendjng Club就要收取滞纳金。发展十分迅速的Lendjng Club，到了2013年1月，贷款发生总额已经超过了12亿美元，中间服务收入也突破了1亿美元。

作为一家典型的P2P网络借贷公司，Lendjng Club的发展情况已经从

某种程度上展示了行业的整个发展趋势。最先发源于英国的P2P网络借贷业务，第一个P2P的创新借贷平台也出现在英国，那便是在短时间内迅速将自己业务拓展到美国、德国等国家的Zopa。紧接着2006年Prosper在美国创立，到2013年1月初，平台的会员已经有160万名左右，会员间的贷款也达成了将近4.44亿美元。与传统银行的金融中介相比，P2P金融交易有着高效率、低成本的特点，服务人群的数量和范围也扩大了，加速了金融发展的进程，只不过双方数据信息的可靠与否是推动该项业务发展的关键所在。

相对于欧美各国，中国的P2P模式起步晚，一直到2007年才真正进入中国，不过发展的速度却十分惊人，很短的时间内像是宜信、拍拍贷、红岭创投、贷帮等一批新兴P2P网络借贷公司纷纷出现，传统的网络大腕也参与到了这个市场的角逐中。中国平安集团于2012年3月投资成立了网络投融资平台陆金所开始运营。几个月后的12月，国家开发银行、江苏省金融办和江苏金农公司合作推出了社会融资服务平台“开鑫贷”。投资机构看到了如此巨大的市场想象空间，也“蠢蠢欲动”，最典型的要数向拍拍贷投资2500万美元的红衫资本。哈佛商学院在2012年5月做了一份《中国人人信贷概览及其在中国金融体系中的潜力》，其中报告提到中国的P2P网络借贷平台数量已经超过了100家，而且还在以每年50%的增长速度在增长。显而易见，因为迎合了市场的需求，P2P市场的发展速度很是惊人。由于不健全的信用体系，传统国内的银行开始纷纷提高自己的贷款门槛，而且在金融危机和欧债危机影响下，银行更是收紧

了银根，原本在资金周转上就有困难的中小企业更是苦不堪言。而对于普通的贷款人来说，股市低迷、楼市受限且理财渠道有限，也受到了不少的影响。此时的P2P借贷平台显然为他们提供了最为有效的资金支持，用一种更为便捷的运行模式让中小企业和闲散资金的拥有者找到了最好的服务平台。

从运行模式来看P2P网络借贷和EBAY有些类似，他们都搭建了一个平台，为借款人和贷款人搭建了服务的网上平台，借款人需要在平台上填写自己的贷款金额、用途、期限、信用记录以及个人信息等资料，通过对这些资料的初步审核，网站会判定借款人的信用评分和风险等级，再为贷款人提供相应的投资方案。竞拍模式是网站的主要交易机制，贷款人利用贷款利率来竞标，最终利率最低的人被选为最终贷款人，成功后网站从中收取手续费。两者的区别就在于P2P网络接待竞拍的是资金，而EBAY则竞拍实物商品。网站为了规避风险，借款人从贷款人那里获得的资金仅为其所需的一部分，贷款人剩下的资金则分散给其他借款人，而借款人所需的其他资金则分散由不同的借款人提供，这样一来，借款和贷款就形成了复杂的交叉。

总体而言，P2P的运行模式同上述的方式很类似，只不过在特殊的中国市场环境中，引入中国的P2P做了一定的调整，衍生出两种主流的发展模式，其中一种的代表是拍拍贷，它们是单纯的线上借贷平台，这和欧美国家的P2P网络借贷模式比较像，这种平台必须依赖完善的市场信用体系和金融服务体系。在中国推行这模式，显然有成本低、扩张速度

快的优势，只不过风险过高。另一种则是线上线下融合的借贷公司，以宜信为代表。这种模式通常线下模式更重，尽管成本高、扩张速度慢，但是风险却低许多。中国的一部分P2P网络借贷平台为了吸引贷款人，会提出“保本承诺”。如果有了违约情况出现，贷款人可以从网站获得本金赔偿，显然这个承诺是为了承担贷款违约的风险，只不过资金链和盈利会受到或多或少的影响。

大数据带来的智能化与柔性化

在商品市场供大于求的时候，企业的价值链就会迅速转向需求驱动。客户在大数据时代介入了企业当中，让企业价值链在引导之下深度整合，推动组织价值的智能化和柔性化，研发、设计、生产、供应、销售、售后服务等等的价值链环节都会有类似的体现。

大数据运用的企业可以对客户进行微分化、个人化的定位，对客户所有的需求特征实时、全面地把握，传统企业显然在智能和柔性程度上缺乏这样的高效。举例来说，Kindle就有这方面特征的功能。用户在Kindle上阅读过的文档都会用附件的形式给亚马逊分配给其的邮箱发送邮件，这个邮箱的真正作用不在于收发邮件，而是Kindle给用户的云存储空间。一旦Kindle连上网络，没下载过的文档就会自动下载下来，云存储空间的任意文档都可以让用户随时随地反复下载阅读。

售后领域，大数据的智能化潜力也是十分巨大的。企业利用远程智能监测、远程辅导等等模式，把人员的投入有效地降低了，服务质量和服务效率更是提高了。例如电梯制造商、飞机制造商、机床制造商等等

设备制造商，会把传感器植入售出的设备当中，随时随地地记录设备的运行情况，设备制造商因此即时获得设备在运行中的情况，对出现的问题进行及时诊断，不需要再到现场，通过远程就可以维护产品。远程智能维护过程中，设备保证系统能够有效运转的核心必须是长期积累的海量数据。

生产流程由于大数据所带来的变化有：为了满足个性化需求而进行柔性生产；为了降低生产风险采用模拟技术；为了改善操作环境而进行实时远程监控。

未来生产商会听到越来越强烈的消费者的声音，消费群体也会因此拥有越来越大的群体力量，消费者最终会成为价值链的第一推动力，所以未来商业模式的主流必然是“定制”。个性化需求必须要求多品种、小批量和快速反应，在竞争当中，能够实现柔性生产的企业会有较大的优势。要实现柔性化生产，成本以及技术是最大的难题，尤其要注意生产中大量应用信息技术，这样才可能降低柔性化生产的成本，适应个性化需求的需要，信息技术的改进还会带来生产线流程的进步，最终实现柔性化生产。实现柔性化的基础在于获取、传输、运用数据，就现在而言这样的应用已经在汽车、家具等制造业中非常常见了。

种种风险都会在生产过程中出现，它们无一例外地会带来巨大损失。首先要对历史经验的数据和规律进行总结，模拟分析生产全过程，预见风险，做到防患于未然。

当人工操作已经和生产环境发生矛盾时，或是人工成本大幅增加

时，人力就会为机器所替代。企业在机器当中植入传感器，实时传回机器工作的数据，实现远程监控设备的目的，操作人员就可以远程操作设备了。

市场需求预测是供应链管理中最为关键的环节。大量个性化的需求可以通过高效分析大数据而获得，加上了解各种辅助信息和合理的预测，供应链中的各个环节都可以灵活、适时地安排。

供应商通常都会累积一系列数据，对需求预测进行改进来安排供应计划。供应链上下游的数据透明了之后，生产和供应的安排就可以变得更为合理了。例如零售商的数据反馈给生产消费品的企业，设备生产商的数据又反馈给生产零部件的企业，在他们获得数据之后，物流、生产、原材料等供应链的环节就可以得到合理调配和安排。此外，上下游之外的数据也可以进行创新整合来创造价值。举个例子，一家跨国饮料企业将每天外部合作伙伴的天气信息合成，再将其反馈给需求和存货规划流程，根据特定日子的温度、降水和日照三个数据点的分析结果，适当减少了欧洲一个重要市场的存货量，这当中预测的准确率也随之提高了5%。

以往企业中的研发、设计、供应、生产、服务、销售等部门的信息总是彼此独立的，随着结构和非结构化数据技术的不断进步，各个环节之间的信息不再是独立的，也是有机地结合在了一起，研发人员不论从哪个环节下手都可以提取到非常有意义的信息，譬如生产新产品中遇到的问题，销售情况，还有客户的反应等等。定期的分析可以让研发人员

获得这些信息，在大数据的时代，研发人员可以实时获得此类数据，通过分析来提取有价值的信息，从而改善产品的设计方案，产品的更新进程和对客户的响应也会随之加速。

在海尔公司内部搭建了市场链研发和营销协同机制，除此以外，参与其中的还有研发部、企划部、市场部和售后部，新产品的市场监测体系建立在多维度的数据之上，一个对市场快速反应的开放系统由此形成，企业为此有了优化产品、推陈出新的重要保障。

长期占据亚洲市场四成份额的UC浏览器，之所以能保持这样的优势，正是因为其先进的研发理念，UC坚持在市场之上为客户创造价值，推出了智能适应屏幕大小、夜间模式和语音等多个适应客户需求的功能，而这一切决策都来自对客户行为数据和反馈的不懈跟踪。UC的CEO俞永福就曾经说过："全球化的进程中，移动互联网企业要做的是全球化思考和本地化执行。在开拓重点区域的市场时，除了要在产品的语言翻译和横向移植上下功夫，还要充分了解当地的文化。"俞永福的意思实际上就是要充分掌握和分析当地客户行为的特征数据。

在电子商务化已经深入企业的买卖关系之后，企业价值链受到大数据的影响就更加显现出来了。阿里巴巴集团的总参谋长曾鸣说过："互联网如果开始继续推动价值链的所有环节，网络和不同的player之间能通过信息实现协同分享的时候，电子商务的威力就会真正发挥出来，而这当中全链条的价值就会得到再造，同时也是个价值创新的过程。"

生活、工作、思维的颠覆重构

大数据带来了时代的重大转型，这就好比是人们可以通过望远镜感受宇宙，通过显微镜观测生物一般的重大发现。人们的生活和理解世界的方式都正在被大数据所改变，一切的新发明和新服务都来自大数据，还有其他更多的改变也在酝酿中。

2009年一种新型的流感病毒出现，那就是结合了禽流感和猪流感特点的甲型H1N1流感，就在几周的时间里这种病毒就快速传播开来。由此全世界的公共卫生机构都在担忧可能会有一场致命流行病蔓延开来。不少评论家开始发出警告称有大规模的流感会暴发，他们担心1918年在西班牙暴发的那场夺走了上千万人性命，且影响了5亿人口的大规模流感的悲剧会再次重演。更为可怕的是，当时没有一个国家已经研发出对抗此类流感的疫苗。公共卫生专家唯一能做到的就是减缓病毒传播的速度，就算是要做到这一点还需要了解的是这病毒究竟在哪里出现。

和其他的国家一样，美国也要求自己的医生一旦发现新型流感病例要尽快报告疾控中心。可是真正患病的人总是在患病多日后才会到医院去，因此真正传到疾控中心的信息总需要一定的时间，正是因为这个流感病例的通报总存在一到两周的延后，疾控中心也只能一周汇总一次数据。可是信息滞后两周对于一种飞速传播的流感来说实在是太可怕了。公共卫生机构因为信息滞后而在暴发流感疫情时总是感觉无所适从。

甲型H1N1流感暴发之前的几周，曾经在《自然》杂志上有一篇互联网巨头谷歌公司工程师发表的论文，很是令人注目。公共卫生官员们和计算机科学家们为之震惊不已。文章中，将谷歌是如何能预测到这个冬

季流感会传播的原因解释得非常清楚，他们认为传播不仅仅是在全美范围内，还会蔓延到特定的地区和州。通过观察人们在网络上的搜索记录，谷歌完成了这样的预测，这用的是一种以前被忽略过的方法。多年来积累的搜索记录都被谷歌给保存了下来，几乎每一天都会收到全世界30亿条以上的搜索指令，这么庞大的数据资源足以让谷歌可以选择用这种方法来推断结果。

谷歌公司拿着美国疾控中心2003-2008年间季节性流感传播的数据和美国人频繁检索的5000万条词条做一个对比，通过人们的搜索记录，谷歌希望以此来判断人们是否患上了流感。除谷歌以外，其他公司也想用其他方式来考证这样的结果，只不过他们不具备谷歌公司那样的庞大数据以及处理、统计数据的技术。

谷歌公司的员工虽然提出了推测，认为在网络上特定的检索词不过是为了得到关于流感的消息，像是“治疗咳嗽和发热的药物有哪一些”，可是这并不是找出这些词的重点，甚至连他们自己也不知道最重要的是哪些词条。关键是他们所构建的系统和这些词义中间并没有依赖关系。这个系统所关注的只有一个，那便是特定检索词条的使用频率和流感时间传播和空间传播的联系。谷歌一共处理了4.5亿个不同的数学模型只为测试这些词条。得出预测的结论之后，再和2007、2008两年美国疾控中心所记录的实际流感病例数据一比对，就发现居然存在着45条检索词条的组合，而在一个特定的数学模型上，他们软件所推测出来的结果和官方数据有着97%的相关。他们居然也能很准确地判断出流感是从哪里

开始传播的，这和疾控中心的判断非常一致，而且还很及时，不至于有延误的情况出现。

2009年甲型H1N1流感暴发的时候，谷歌显然要比习惯性滞后的官方数据要更为有效，更为及时。公共卫生机构的官员因此得到了最有价值的数据信息。谷歌公司做到这一点居然不是通过分发口腔试纸或是联系医生，而是建立在与医学毫无关联的大数据的推测上。大数据可以称作是现代社会所具备的一种新型能力，用一种前所未有的方式分析海量数据，从而产生巨大的价值或是深刻的见解。可以想见下一次流感来袭时，这样的技术理念和数据储备会让全球人们拥有更好的预测工具，防止流感的大面积蔓延。

不但是公共卫生领域正在被大数据改变，几乎整个商业领域都在因为大数据而重新整合，譬如航空领域。

2003年，正准备从西雅图飞往洛杉矶去参加弟弟婚礼的奥伦·埃齐奥尼（Oren Etzioni）开始在网上订购机票，他知道越早买价格越便宜，因此几个月前他就在网上订购了一张机票。上了飞机，奥伦·埃齐奥尼非常好奇地向其他乘客询问购买的价格，当他听到一个比他晚买很多天却价格便宜不少的乘客的话时，感到非常气愤气氛，结果越问就越是发现，很多人的票买得比他便宜。

大多数人或许碰到这样的事情很可能在他走下飞机的那一刻就会忘掉。可是作为美国最知名计算机专家之一的奥伦·埃齐奥尼，从他第一天开始担任华盛顿大学人工智能项目负责人开始，他就创立了很多典型

的大数据公司，即便大数据这个概念在那时候还没有兴起。

1994年，在埃齐奥尼的协助下，最早的网络搜索引擎Meta Crawler诞生，后来为Info Space公司所收购。埃齐奥尼还联合创立了第一个比价网站Netbot，随后又被Excite公司所收购。还有路透社收购了他创建的能够在文本中挖掘信息的Clearforest公司。在埃齐奥尼看来整个世界其实就是一系列的大数据，而自己可以来解决当中所有的问题。埃齐奥尼1986年毕业于哈佛大学，作为第一届计算机科学专业的毕业生，他所有的精力都花在了如何解决这些问题上。

当飞机着陆，埃齐奥尼决定要开始开发一个新的系统，作为预测网络机票价格的系统。同一航班上的同一座位本来应该有相同的价格，但事实上人们购买它的价格却千差万别，而这当中的理由只有航空公司心里清楚。

埃齐奥尼认为，机票价格的奥秘无需他来解开，他所要做到的就是帮助人们预测未来一段时间机票价格会跌还是会涨，这个想法是可操作的，只不过操作起来还存在一定的困难。很显然，这个系统为人们所展示的是与购买时间无关的特定航线机票的销售价格变化。

一张机票的价格若是呈现下降趋势的话，系统就会提醒顾客可以延缓购票的时间，反过来如果是上升的趋势的话，系统就会提醒用户当下就要买下机票。总而言之，埃齐奥尼开发了一套针对9000米高空的加强版信息预测系统，事实上这是个庞大的计算机科学项目，而它却是十分可行的。于是埃齐奥尼着手启动了这个项目。

埃齐奥尼所创立的预测系统，虚拟的乘客可以通过它节省不少经费。41天内12000个价格样本是这个预测系统的数据基础，而这些数据均来自一个旅游网站。预测系统能做到的是预测会发生什么，至于原因它是无法说明的。换句话说，它所推测的是推动机票价格波动的因素。机票降价的因素很可能是因为季节原因或是卖不出去等等，还可能是由于“周六晚上不出门”等原因，这些和预测系统都没有关系，它不过是通过数据的分析和统计来预计未来的趋势。埃齐奥尼总在思考是买还是不买这个问题，为此他还给自己的这个研究项目取了一个非常贴切的名字——哈姆雷特。

这个小小的项目逐渐壮大，成立了一家科技创业公司，还得到了风险投资基金的支持，取名为Farecast。该公司利用对机票价格走势和降价幅度的预测，指导消费者不错过最佳的购买时机，而这些都是消费者此前在其他网站获得不了的信息。

从保障自身透明度的角度来考虑，机票价格走势的预测可信度在这个网站上都可以查到，消费者可以以此进行参考。而海量的数据正是支撑这个系统运转的基础。为了能让预测更为精准，埃齐奥尼发现了一个行业机票预订数据库。基于美国商业航空产业数据而推测出的结果，几乎每一条航线上的每一个航班的每一个座位一整年的综合票价记录就是推测的依据。到现在，Farecast所拥有的飞行数据记录已经超过了2000亿条，用这种方式，消费者利用Farecast省下了不少开支。

埃齐奥尼有着棕色的头发，笑起来常常露出牙齿，无邪的面容，怎

么看都不像是那种有魄力让所有航空公司一年损失数百万潜在收入的人。可是他确实是个目光非常长远的人。埃齐奥尼2008年开始在其他领域应用这项技术，像是预订酒店、二手车买卖等等。尽管所有领域中的差异不大，但是还是有大幅度的价格差存在，并据此有很多可运用的数据都是这项技术的基础。埃齐奥尼在实现自己计划前，微软公司就与之协商以1.1亿美元收购了他的公司。此后，Farecast并入了微软的搜索引擎中。

Farecast不过是所有大数据公司的一个小小的代表而已，它呈现出的是当下世界发展的趋势。时间往前推五年到十年，埃齐奥尼绝对想不到这个。他自己也说过这仿佛是不可能的，只因为那时的计算机处理和存储价格太贵了。这一切发生仰仗技术上的突破，但这并非唯一的原因，因为还有很多细微的重要改变在发生，尤其是人们有了数据使用的理念。

现在不再有人认定数据是静止的，不过从前收集了数据之后人们就会认定该数据没有价值了，就比方说航班落地之后，票价数据也就用处不大了（对于谷歌来说，检索命令完成以后数据就没有用了）。

因为数据而带来的便利随处可见，现代人人手一部手机，办公桌上均有一台电脑，而在办公室和办公室之间还有大型的局域网相连。不过，似乎人们还不是太重视信息数据本身的用处。计算机技术在过去的半个世纪已经融入了社会生活的每个角落，信息大爆炸的时代到来还引发了重大的变革。世界因此有了比以往更多的信息，增长速度也随之加

快。信息总量的变革引起了信息形态的变化，换言之就是量变带来了质变。其中首当其冲的如天文学和基因学，经历了信息爆炸，并因此提出了大数据这个概念。现在所有的领域几乎都有了这个概念的应用。

大数据这个概念并不确切。起初大信息量的处理就是大数据，它一般指的是超过一般电脑处理数据的常用内存量，所以工程师为大数据而改进了数据处理的工具，新的技术处理方式也因此诞生，譬如有谷歌的MapReduce和开源Hadoop平台（最初源于雅虎）。人们有了这些技术，能够处理的数据量飞速增加。更为重要的是，数据的排列也不再需要依据传统的数据库表格，很多消除僵化层次结构和一致性的技术也在这个时候应运而生。与此同时，互联网公司还因此收集了大量充满价值的数据，并且在这些数据利用上发现了强烈的利益驱动力，而互联网公司也成了信息数据处理技术的领头羊，这一切很顺理成章。它们的发展让它们成了新技术的领衔者，甚至超过了有几十年经验的线下公司的发展。

现在最可能的一种方式就是人们认定以大数据为基础能做到很多事情，而这一切在以往的小数据时代是几乎完成不了的。人们通过大数据来获知新知识，且创造新的价值，同时大数据也改变了市场和组织机构，包括政府和公民的关系。

第七章 得数据者得天下，商业竞争中的大数据

商业领域无疑是大数据时代的领头羊，它们最先发现了大数据在商业竞争中的价值。数据化的进程推动了商业数据的可量化变革，同时也更新了对客户形象的描述。信息化的商业竞争已经打响，谁掌握大数据谁就会是未来的赢家。

大型公司的垂直一体化趋势

近几年，信息产业的垂直整合愈演愈烈。IBM作为计算机的缔造者之一，自始至终都为客户提供包括存储、主机、操作系统、数据库、中间件、应用软件等等在内的完整解决方案，可以称作十足的蓝色巨人。而另一家由拉里·埃里森领衔的甲骨文（Oracle）公司，第一步就是占据了数据库的软件市场，再以此向应用软件的市场扩张，它在全球企业管理软件上的业务仅仅在SAP之下。甲骨文似乎还不满足，还在继续收购，利用自己庞大的客户群逐渐向产业链的上游进发，第一步就是收购了一家UNIX主机厂商——SUN公司。还有家喻户晓的微软集团，先是在操作系统上坐稳第一把交椅，随后就开始向产业链的下游扩张自己的业务，很快推出了自己的数据库产品，并购了众多小型应用软件公司，还为企业提供管理服务。

这几年海外的这些数据巨头中发展最具代表性的就是甲骨文公司。而在国内，华为则是垂直整合信息产业链中的代表，它每走一小步都是国内整个信息产业发展一大步的标志。在德国，软件巨头SAP，也是用友这家国内软件公司的最大竞争对手，收购了一家数据库公司后，就开始坚实地向上游垂直整合迈进。在这一浪潮中，用友软件要如何更好地赢得市场竞争，产业界和资本市场都十分关注。或许，用友也会走甲骨文和SAP的道路，同样进军产业的上游。

埃里森让甲骨文公司变得非常具有进攻性，在开放平台数据库上，甲骨文已经是市场中的佼佼者，随后就朝下游扩张，向整个软件领域铺

开。2004年甲骨文收购了Peoplesoft，这原本是其竞争对手，同时也是一家企业人力资源管理软件的厂商；2005年它又收购了Siebel这家全球最大的CRM软件厂商，很快就成了全球第一的CRM应用软件供应商；2007年它收购了Hyperion（海波龙）这家商业智能分析厂商，目的是为了向客户提供应用软件、咨询服务，以此来加强终端客户的掌握，最终成了与SAP平起平坐的企业管理软件供应商；2008年，它收购了Primavera软件公司，开始了项目组合和管理软件的生产供应，一年后项目管理软件产品升级时，正式更名为Oracle Primavera。

甲骨文公司的下一步就是朝产业链的上游扩张，全方位地增强服务能力。2008年，甲骨文收购了中间件巨头BEA，开始了自己基础软件领域的发展。它的举动让中间件市场发生了翻天覆地的变化，使得众多其他的小中间件厂商难以维持。甲骨文随后又开始走上了硬件的发展道路，2009年甲骨文收购了SUN，借此和自己的系统创造极佳的互补性。当时的SUN有SPARC处理器和Solaris操作系统，也就在那一年，虚拟化产品商Virtuallron的收购工作也完成了。甲骨文因为收购也补齐了自己的短板，变成一家同IBM一样的业务涉及所有硬件和应用领域的IT巨头。一时间关于“红色巨人PK蓝色巨人”的讨论兴起。甲骨文的垂直整合在收购工作完成之后，股价一路攀升，居高不下。

甲骨文收购了SUN后即刻发布了一款新型的数据库云服务器ExaData，它集SUN软件和甲骨文软件的传统优势，主要负责数据仓库和商务智能类系统、OLTP类系统、混合负载类系统、数据库云平台服

务。这当中服务器、存储、IO和虚拟化软件都被甲骨文集成在了一起，体现了它对数据库、中间件和应用软件的理解。

2012年甲骨文公司的第三季度财报显示，这个财季因为推出了软硬一体化的集成设计系统，硬件收入增长了139%，无疑成了甲骨文历史上成长速度最快的产品。

国内垂直整合思路最清晰、坚决的就是华为公司。华为一度将IBM视为自己奋斗的目标，但现在IBM却是华为最大的潜在威胁。相比于IBM，华为的产业链更多出了一层“网络设备”。虽说IBM的应用软件更为完备，可是华为实施了“被集成”战略，国内多家应用软件商与之合作后，华为集成了产业上下游的产品，客户也因此获得了最为完整的解决方案。

如华为一般的大公司，一定会从产业战略的层面开始思考公司的安全问题，还有产业链上下游的哪些公司是否扼住了自己咽喉的事情也必须考虑。说起来道理都不复杂，华为对对手的威胁已经存在了，那么对手们自然也要从操作系统和CPU等等方面扼制住华为的发展。就此而言，大型公司必然会走上垂直一体化的道路。

客户形象的丰富源于对客户的全面理解

理解客户是客户为中心的起点，理解客户最有效的途径就是大数据，大数据无论是划分客户、定义客户还是分析客户需求都有着不可比拟的精准高效性能。

先说一下划分客户，大数据能把客户划分得更为微分化、个人化。

通常把传统市场划分为大众市场、利基市场、细分市场几类。大众市场当中，会有一个大范围的客户群，其中涵盖了企业的产品、服务、渠道和营销推广，客户在这个客户群中的需求基本相同，好比是个人计算机或是运营商的通信服务。利基市场当中则是针对某一个特定市场进行专门定制，包括了产品、服务、渠道和营销推广各个领域，像是专门提供汽车零部件等等。接下来是细分市场，产品、服务、渠道和营销推广则是针对不同需求的市场群体区别对待，像是SMH有着不同品牌的手表，宝洁则是针对不同需求推出三大品牌的洗发水。

在大数据时代，市场的划分不再沿用这样的概念，而是进入了微分化，个人化，这要归功于以下两个条件。

第一个条件是丰富的数据量，市场微分化和个人化必须依赖这样的信息基础。传统行业除电信、银行之外，极少有机会获得单个客户的详尽数据。客户的数据随着大数据的到来而呈现爆炸性增长。内部信息开放的通道，也就是“联网型组织”利用互联网让客户和供应商们参入进来，而它们本身就优先有了这方面的基础，像是淘宝、京东商城、腾讯、百度、新浪等等企业。除此之外，还有如奔驰、海尔、苏宁电器这样的传统企业也在朝“联网型组织”发展。

第二个条件则是持续进步的大数据技术，微分化、个人化的市场因此而有了发现价值的可能。在亚马逊上亿客户的单体客户行为数据都被收集了，比如客户搜索了什么，对哪些产品的详细介绍有兴趣，最后买了什么产品等等都是亚马逊记录下来的数据。其他用户的购买数据对个

人客户而言也可以成为推荐。很显然，很多个体客户在购买时，更愿意去看看和自己爱好相似的人搜索了什么，买了什么，做了什么评价。

再来说说客户定义。有了大数据，标准化的传统客户定义也转化为个性化。

传统的客户定义通常依照客户价值来判断，或是用问卷访问或是用小组访谈这样的调研技术对客户进行分析，再找出一个或多个维度来细分定义市场，给予客户群一个标准化的面孔，配以企业本身的产品、服务、渠道和营销推广。招商市场利用银行“一卡通”客户的总资产量，客户因此被分为普通卡、金卡和金葵花卡VIP三类客户。还有中国移动也将自己的客户分为追求高端的全球通客户，年轻动感的动感地带客户以及性价比极高的神州行客户，这一分类也是依据客户价值、客户行为和需求划分的。

大数据时代的客户定义不再延续群体标准化的划分标准，而是全面、立体地描绘客户形象。两方面的数据组成了客户形象，一个是结构化的交易数据，其中有消费水平、消费频次和生命周期等等，第二个是非结构化的交互数据，譬如文本、图片和多媒体等等。无论是哪一种数据的增长都是远远大于交易数据增长速度的。交易数据和交互数据在大数据技术的综合分析之下，客户的形象就越来越立体和丰满起来，不再是传统的标准和抽象的面孔，这当中也将客户的需求精准地反映出来。

还有一部分零售企业在大数据的利用上都走在了前列，运用“情感分析”的技巧将社交媒介上的消费者所产生的海量数据进行挖掘，客户的

情感变动也能及时掌握。企业有了这些数据分析的结果更能即时调整产品和推广策略，商品的周转速度和毛利空间也因此提升。

在大数据定义客户方面，最先有实践的是谷歌公司。他们理解客户行为和习惯的途径是免费软件和服务，提供给客户的软件越多，就越是能收集、理解客户，像谷歌推出的谷歌图片、谷歌音乐、谷歌邮箱、谷歌视频等等软件均为谷歌从不同方面了解客户提供了可能。在精准地获取客户信息的基础上，谷歌还可以获得精确的广告服务信息，高利润的商业模式因此诞生。

最后来说说实时需求。客户需求信息当中最为有效的是客户实时行为倾向，只不过这种信息是瞬时的，传统方法要捕捉是很困难的。而在大数据时代，客户实时的个性化需求都能够被获取，企业能针对这个做出高效的决策。

零售行业的传统其实只要稍加改进就能实时获取客户的行为信息，像是在购物车上安装传感器，客户的行进路线就会被实时追踪，例如不同位置的停留时间，还有最终购买的物品数量品种等等。这信息对卖场在货架展陈、商品上架的调整方面十分有帮助，利润率和销售额也有所提高。

互联网零售企业除了能获取实时信息，对于用户行为的分析、公司营销策略的调整和分析也可以从中获益。互联网点击率反映的就是客户的行为和偏好，可以根据此类数据对客户行为进行建模，并由此向客户推荐优选商品，推出省钱的奖励计划，那么整个销售流程就非常圆满

了。

线下的商家更可以根据实时的数据反馈来向周边客户推广优惠活动。如今智能手机的普及让定位于手机位置信息的应用开始快速发展。只要一个携带智能手机的用户进入运动服装店，服装店就可以利用这个智能手机向其他周边的用户来推送最新的优惠活动信息，从而提高销量。社交客户端以此位置信息为基础也可以来做此类的推广，像是已有商家在微信上做此类的工作了。

量化奠定了数据化的内核

原始社会和先进社会的分界线之一就是记录信息的能力。最早期文明中古老的抽象工具源于基础的计算，包括计量长度和重量。早在公元3000年之前，在印度河流域、埃及和美索不达米亚平原地区，信息记录的水平就比较高了，日常计量也有十足的长进，尤其是在美索不达米亚平原一带，产生了一种记录生产和交易的精确方法，这是书写进步的功劳，早期文明也因此得以计量和记录下来，为日后的研究提供了依据。因为有了计量和记录，也就有了数据，这是最早的数据化根基。

人类活动在计量和记录中得到重现，像是建筑物的建筑方式和使用的原材料被记录下来，从此后就可以用类似的方法来建造房子，或是进行实验等等，譬如以改进方式来建造新的建筑物，然后依次循环。一旦有了交易记录的话，那么一块田地丰产时的稻谷产量是多少就会很明确，还有政府税收多少也能因此了解。预测和计划必须有计量和记录为基础。记录让交易的双方明确了对方的赊账状况，要是缺失了这些凭证

的支持，那么欠债的一方绝对可以赖账。

计量在几百年当中从最初的长度和重量领域慢慢扩展到了面积、体积和时间之上。公元前的最后十个世纪，西方的计量方法基本都已经产生，尽管它还有众多的缺陷存在。早期文明，计量方法在计算上总有很大的困难，即便是非常简单的计算，像罗马数字的计算系统在数字计算上就不太合适，原因在于记数制上缺少十进制，哪怕是专家碰上大数目的乘除都会感到困难重重，而一般人通常就是简单的乘除都无法完成。

印度在公元1世纪的时候就发明了一套数字系统，随后几经辗转到了波斯，波斯人对其进行改进，又传到了阿拉伯国家，阿拉伯人大大改进了它们，它们就成了现在人们广泛使用的阿拉伯数字的前身。十字军东征尽管使当地的人民陷入战火之中，可同时又让地中海东部迎来了西欧文明，其中最重要的一个部分就是阿拉伯数字的引入。在公元1000年，教皇西尔维斯特二世在民众中推行阿拉伯数字。到了12世纪，阿拉伯数字的介绍书籍被翻译成了拉丁文，在整个欧洲地区传播，算术因此迎来了自己的春天。

阿拉伯数字在传到欧洲之前，算术就因为计数板的使用而有了较大的改进。所谓计数板就是在光滑的托盘上用代币来代表数字，代币的移动就可以被用来表示某个区域的加减。只可惜计数板也有自己的短板，过大或是过小的计算是不能同时运算的。还有一个缺陷是计数板上由于数字变化过快，只要是微小的一个碰撞或是错误的移位都可能会导致南辕北辙的结果。显然计数板尽管可以用来计算，但它不适合记录。数字

记录一旦在计数板之外的话，计数板上的数字还是要转换成罗马数字，这无疑太费时费力了。

算术因此给数据添加了新的价值，除了可以被记录下来以外，还可以再分析和利用。12世纪开始阿拉伯数字就出现在欧洲，可是到16世纪晚期才真正被广泛使用。16世纪开始数学家就开始向人们宣传阿拉伯数字的优势，相比计数板计算的速度要快6倍。可是真正让阿拉伯数字为大家所接受还是由于复式记账法的出现，显然这就是数据化工具中的一种。

会计手稿早在公元前3000年就已经出现了。只不过在接下来的几百年里它始终缓慢发展，而且大多数时候都停留在了记录某个地方特定的一场交易的阶段。当时对于记账人和他的雇主来说，最重要的莫过于去判断一个账户或是自己从事的某个行业是否能盈利，记账法很明显已经无法轻易地得出答案。14世纪的时候，意大利会计已经开始流行用两本账本来记录交易，这种尴尬的局面也就得到了挽回。通常这样的记账法优势是借贷相加，就能够通过表格来了解账户的盈利和亏损情况。从那时起数据就开始有了它的利用价值，只是它仅仅是局限在读出盈亏情况罢了。

现在的会计业和金融业不断发展，复式记账法被视为其不断发展的结果。实际上在推进数据利用的过程当中，这就好比是一个重要的里程碑。因为相关的账户信息的分类记录由于它的出现而实现。它的基础是一系列记录数据，也是信息记录标准化最早的例子，会计们因此能互相

看懂对方的账本。复式记账法的规律能让账户盈亏情况的查询变得更为简单，交易的记账线索能够从中发现，这样一来所需的数据就越容易找到。在设计理念中它包容了“纠错”的思想，今天的技术人才可以以此为借鉴。要是有一个账本中有问题，那么就可以从另一本账本中进行查询。

复式记账法和阿拉伯数字一样也并非一开始就受到欢迎。200年以后，它广受欢迎是源于一个数学家和一个商业家族的推荐，也是他们改变了数据化的历史。

说到这名数学家，他就是知名的修士路萨·帕西奥利（Luca Pacioli）。1494年，他有一本推荐给普通读者和商人的数学教材出版。此书一经出版大获成功，曾经风靡一时。这本书是历史上第一本全部使用阿拉伯数字的书籍，也因此推动了阿拉伯数字在欧洲的广泛使用。此外这本书的最大意义在于详细论述了复式记账法。此后的几十年，此书关于复式记账法的论述被译成了6种语言，在世界上有多个通用版本。

而上文提到的商业家族，其实就是美第奇家族，他们是威尼斯商人和艺术资助人。16世纪欧洲最有影响力的银行家族非他们莫属，这是由于他们采用了当时世界上最为先进的数据记录方法——复式记账法。由于帕西奥利的著作和美第奇家族的成功，也让复式记账法一举成为标准数据记录法的基础，当然也让阿拉伯数字成了不可替代的数字。

数据记录不断发展，人们对世界的探寻也在不断前进，似乎关于时间、距离、地点、体积和重量的精准记录成了所有人的愿望。19世纪，

科学家们陆陆续续发明了各种记录和测量电流、气压、温度和声频的单位和方式，科学一下子进入了量化的时代。好像不论是什么事物都可以被测量、划分和记录，人们对单位量化和记录的疯狂程度甚至已经到了需要用精细的划分来理解人的智慧了。不过这一轮伪科学的热情很快就消散了，而人们对一切量化的工作热情似乎从未消退。

测量数据和记录数据因为有了新工具和开放的思维得到了蓬勃的发展，因此也催生了现代数据化。所有数据化的根基都已经搭好了，这一切在模拟时代是不可能完成的，还可能需要无止尽的激情和耐心，不夸张地说，怎么也要有奉献一生的准备，这其中最典型的例子就是16世纪就没日没夜地观察天体运动的第谷·布拉赫（Tycho Brahe）。模拟时代成功的数据化例子并不多见，因为缺乏很好的运气，需要将大量的偶然事件叠在一起。最为幸运的应该是中校莫里，因为负伤他只好坐在办公室，也因此发现了珍贵的航海日志，只是这份幸运并非人人都有。在数据化的实现过程中必须有很重要的一点，也就是挖掘出潜藏在数据中的巨大价值，这才能揭示出最新的奥秘。

数字测量和存储设备因为计算机的出现而应运而生，数据化的效率也提高了不少。计算机的数学分析从数据中挖掘出了越来越多的潜在价值。简单地说，数据化是数字化的结果，但是数字化最终是替代不了数据化的。模拟数据成为计算机可读的数据是源于数字化，但是它和数据化之间的区别是本质化的不同。

文字的数据化进程

那么数字化和数据化有什么不同呢？这个问题不难回答，想要真正理解就来看看两者同时存在且一起起作用的领域就可以了，而这个领域也不特别，就是书籍。谷歌在2004年发布了一个充满野心的计划——数字化所有拥有版权的书籍内容，这样的结果就是世界上所有人都可以通过它们的平台来免费阅读这些书籍。谷歌为了完成这个庞大的计划，和世界上最大、最知名的图书馆合作，并在此基础上研发了自动翻页的扫描仪，做完这一系列准备工作，谷歌省去了扫描上百万书籍工作的大部分费用。

谷歌第一步要做的就是数字化文本，它扫描每一页图书，然后存进谷歌服务器中的一个分辨率很高的数字图像文件当中。这样一来，书本上的文本内容就数字化了，不管是什么地方的人都可以通过电子平台来阅读。只不过用户必须很清楚自己想要阅读的内容在哪本书上，或者可以轻松地从浩瀚的内容中寻找到自己所需要的片段。很显然这是由于文本没有被数字化的缘故，很难通过搜索来检索，更不可能被分析。谷歌所做的是将文本图像化，人的阅读才能让这些图像成为有意义的信息。

这是一个比历史上任何一个图书馆信息量都要庞大的现代的、数字化的图书馆，可是谷歌要做的不止于此。谷歌明白，数据化是下一步必须实现的，只有这样才可以挖掘出这现代化图书馆的巨大价值。从那以后谷歌开始利用识别数字图像的光学字符识别软件来对每一个字、每一句话、每一段进行识别。经过这个过程，数字化文本就转化为了数据化文本了。

现在人类检索这些文本成为可能，而计算机处理和分析这些文本的可能也因此实现了。在检索和查询的过程中，我们能做的还有无穷无尽的分析。通过它们还可以揭示出第一个词或是词组第一次出现的时间和它成为流行词的时间，几百年间人类思维发展和传播的轨迹也可以从中瞥见一斑，好几种语言的分析都可以从中实现。

15世纪中叶，印刷机问世，在这之后大概有1.3亿册的图书出版。截止到2010年，也就是在谷歌数字化图书计划执行的第七年，有2000万册的图书被扫描成了数字图书，大约是人类所有书写文明的15%已经数字化了，这个数字实在太惊人了！这也催生了一个新的学术研究方向——文化组学。这是一个计算机领域的专业词汇，实际上就是把人类的行为和文化发展趋势用文本定量分析的方式揭示出来。

哈佛大学的一项研究利用了几百万册图书和超过5000亿个单词来深入研究，发现这当中有一半出现在这些书籍中的单词是字典中未曾收入的。于是他们在分析报告中提出正规词典中尚未收入的不规范词汇数量十分惊人，这是个巨大的宝藏。在系统分析之下，人们提及纳粹德国时期的犹太画家马克·夏加尔（Marc Chagall）时，研究人员发现很多时候个人或是思想受到审查和压制，有很多“可量化的痕迹”被流传了下来。词语之于书籍仿佛是化石之于岩石，凡是信奉“文化组学”的人都和考古学家一般，很多宝藏都会被一一挖掘出来。至于其中的偏差也是必然存在的，就像是图书馆的每一本书难道都反映了最真实的现实吗？事实上难道不是图书馆员或是作者自己所看到的现实吗？就算是这样，人们还

是可以通过“文化组学”来获取极大的惊喜。

文字如果成为数据，它一下子就有了巨大的能量，人们可以阅读，机器还能分析。谷歌作为一家典型的大数据公司显然明白信息化和数据化的价值，更何况很多数据的潜在用途是非常多的。因此，谷歌开始利用数据化的文本对自身的翻译系统服务进行改进。这个系统本身是可以自动扫描译本的，随后从两种语言的匹配中找到对应的词或是词组。一旦有了答案，系统就会用类似数学方法的方法来对待翻译，随后在电脑上找到两者之间最为恰当的词或是词组。

谷歌并非这一领域第一个吃螃蟹的人，也不是计算机时代第一个利用数据化来唤醒文明财富的公司。早在1971年，有一个志愿者就提出能否将公共领域的书籍数字化，制成电子书放到网络上，以供更多的人来阅读，这就是著名的古登堡计划（Project Gutenburg）。这是个非常有意义的举动，尽管当时还未能将书籍数据化，而书籍的其他功能也尚未开发，毕竟他所提出的是要扩充书籍的用途，这是关注阅读的一种做法。同样出版社这么多年也将电子书籍的开发视为未来的战略考虑，不过他们的重心还都没有把数据纳入自己的商业模式，显然书籍内容才是他们所关注的核心价值。所以书籍的数据价值始终没有被发现，只因为他们没有发现数据化的需求，更不知道书籍数据化的潜力。

如今，不少在电子书领域竞争的公司，譬如亚马逊推出了它的kindle电子书阅读器，一跃成了这一行业的先驱，从中可以发现亚马逊和谷歌之间差异性的发展策略。

亚马逊的资源来自数据化的书籍，而对于数据化背后的价值却从未关注。作为创始人的贝索斯说服了上百家的出版社在自己的kindle上发布图书，而读者在kindle上所读到的书籍并不只是简单的数字图像，而是可以更改字体和颜色来读的数据化文本。亚马逊数据化的文本已经超过了上百万册，可相比之下谷歌在做的却是费力地把所有旧的版本来数据化。因此亚马逊的着眼点在于书籍内容的阅读，而非数据化文本的分析。不可否认的是，传统出版社的压力或者限制了书籍内容的应用，毕竟人们还掌握着版权。对于谷歌来说，一个喜欢离经叛道的大数据公司不会有这样的压力，因为它所有的资源都是源于客户的点击，和出版社没有关联。至少现在可以公正地说，亚马逊在数字化内容意义方面最为老练，而真正触及数据化内容价值的是谷歌。

地理位置的数据化构建

世界上最基础的信息就是地球本身。可是从古至今没有人将它完完整整地量化过和数据化过。事实上这些信息的组成部分就是人和事物的地理定位，否则古人怎么会有“所谓伊人，在水一方”的诗句，只不过数据化是这些信息现在面临的问题。

西方的希腊是量化位置信息最早的地方。早在公元前200年，埃拉托色尼就创造了划分区域的格网线系统，这就类似经纬度法。不过这一系统很快就退出了历史舞台，和古代的很多好想法一样。过了1500年，也就是1400年，一本由托勒密撰写的《地理学》的影印本一路辗转从君士坦丁堡到了佛罗伦萨，那是一个科学和古典知识的热点被文艺复兴和贸

易船只点燃的时代。这本著作一下子轰动了整个欧洲，这套系统直到现在仍然被用在解决航海导航的难题之上。自此地图上出现了经纬度和比例尺。1570年佛兰德制图师墨卡托改良了这套系统，从此海员出海时就可以通过它来完成最为笔直的航线了。

尽管记录地理方位的方式当时已经出现了，可是广为认可的标记标准却还未问世，信息共享还有非常大的难度。人们都在呼唤一套标准的标记系统，这和后来的人们呼唤互联网的域名一样。经过了一个漫长的时间经纬度的标准化才出现。1884年，国际子午线会议在美国华盛顿召开，与会的25个国家中有24个都同意将本初子午线和零度经线穿过的地方规定为英格兰的格林尼治。到了20世纪40年代，用墨卡托方位法世界被完全分为了60个区域，世界地理定位的精确度因此提高。

有了这个所有的地理定位信息，也就有了标准化数值范式的标记、记录、测量、分析和共享。可是人们却很少这么做，原因是在模拟数据时代，所有测量和记录地理位置信息的费用都太高了，这也奠定了人们呼吁低成本地理方位测量工具出现的基础。直到20世纪70年代，还必须依靠地标、天文星座、航位测量法和还未成熟的无线电定位技术来为地理位置定位服务。

1978年发生了一场巨大的转变，简称为GPS的全球定位系统，总共24颗卫星首次成功升天。地面上的汽车导航系统、智能手机以及其他的接收器都能接收到它的信号，再利用接受时间上的差异来完成三角定位，而这些信息均源于离地面20372米的高空。这个系统到了上个世纪

80年代首次开放民用，90年代正式投入使用，还开放了商用。十年之后GPS的精确率大幅提升。今天，地理定位系统已经完成了精确到米的定位，也因此实现了古今中外航海家、制图家和数学家的梦想。在和技术手段结合之后，定位系统的发展非常迅速，同样也降低了自身的成本，而这一切都不需要依赖任何专业知识。

每时每刻定位都提供了众多的信息。可能的话，埃拉托色尼或者墨卡托可以定位所有他们所处的位置，这不依赖任何人。不过假设这是可行的话，也不够现实。接收器早期的成本是很高的，技术也很复杂，它只适用于潜艇而不是出租车。幸运的是，数字设备中廉价芯片的普及让一切都发生了变化。20世纪90年代GPS的价格还高达几百美元，现在已经降至1美元以下。如今的GPS的定位仅仅需要几秒钟的时间，它的方法是标准化的坐标表示法，像是37°14′06″N　115°48′40″W，就是内华达州偏远的51号区域，那里有一个美军的高级保密的军事基地，据称那里的秘密就是有很多外星人。

GPS现在是众多定位系统中的一种，包括中国和欧洲国家在内的众多国家都开始了自主产权卫星定位系统的开发工作。不论是哪种系统都要依赖电塔和无线路由器的信度来确定地理位置，所使用的技术仍然是三角测量，这从一定程度上填补了GPS在室内和高楼中缺少定位的缺陷，这是谷歌、苹果等众多公司纷纷要用开发定位系统来辅助GPS的原因。例如谷歌就利用街景车（Street View Cars）一边拍照，一边来收集无线路由器的信号，而苹果的IPHONE手机本身就是一个移动的信号接

收器，在不知不觉的情况下就会收集用户的位置和无线数据，再传回苹果公司。此外谷歌的安卓系统和微软的手机系统也在做同样的工作。

人和事物的地理位置同样都可以被定位。汽车安装上无线传感器，地理位置的数据化也会因此带来保险概念的深刻变革。数据当中包含了时间、地点和实际行驶路程的所有信息，保险公司可以依据这些来为车险定价。英国的车主买保险就是依据他自身的驾驶地点和时间，这些和他的年纪、性别和履历关系都不大。这种保险定价的方式能激发人们更好的行为习惯。与此同时保险的基础也发生了改变，从前考虑的是一个群体的风险，而现在则是个性化的风险分析。汽车定位了每个人的地理信息使得某些固定资产投入的模式也有了改变，例如适用公路和其他基础设施的司机和其他人也能够因此分担其中的部分投入。当然上面所说的这些在所有人和事都以数据形式保持持续定位之前还是很难实现的，而这是未来的一种趋势。

莱维斯曾经说过："预测给我们的是知识，而知识给予我们的是智慧和洞见。"他相信这个系统在用户意识到问题之前一定会预测且解决问题的。

人类身上的数据化实时位置信息运用最为突出。这么多年来无线运营商为了提升移动互联网的服务水平总在收集和分析这些信息。而这些数据的应用范围越来越广泛，很多第三方的新服务也开始倚靠这些数据了，就比如说智能手机不论是否有定位功能，但无一例外地在收集此类信息，另外一部分应用程序的存在目的就在于获得用户的位置信息，像

是Foursquare，用户可以在最喜欢的地方“check in”，然后经过忠诚度计划、酒店推荐还有“check in”地点附近的各类推荐来得到好处。

收集用户的地理位置数据的能力无疑已经成了最为有价值的能力，即便是个人层面，他居住的地方和他想要去的地方这些数据都可以用来预测要如何定制广告，何况聚集起这些信息还能发掘事物发展的趋势。例如大量的位置数据可以为公司预测交通情况提供帮助，或许这一点很难想象，高速公路上的手机居然替代了汽车数量和移动速度来预测了相关的趋势。Air Sage每天将收集来的手机用户的150亿条位置信息进行处理，为超过100个美国城市提供各类交通信息。而Sense Networks和Skyhook两个位置数据服务商则是利用地理位置数据来推测每天晚上最繁华或是最热闹的地方聚集了多少人。

只是在商业之外未知数据的应用价值似乎更大。麻省理工学院的媒体实验室人类动力学实验室主任亚历山大·彭特兰（Alexander “Sandy”Pentland）和他的学生南森·伊格尔（Nathan Eagle）就是最先进行“现实挖掘”研究的学者。所谓“现实挖掘”其实就是在收集大量手机数据基础上进行处理，以此来发现和预测人类的行为。他们的一项研究分析了每个人去了哪里，见了谁，还将易于感染流感的人群给区分出来了，这种区分甚至是在易于感染的患者尚未感染之前就做出来了。一旦暴发了严重的流感疫情，无数人的生命就会因此被挽救，很显然要隔离哪些人，到哪里去找到他们是很容易做到的。不过这些数据若是落入坏人手中，那会有什么样的后果难以想象。

无线数据科技公司Jana的创始人是伊格尔，他收集了100多个国家，超过了200个无线运营商的手机收据，这当中有拉丁美洲、非洲、欧洲在内的35亿人口。伊格尔研究的问题很简单，是关于每一周家庭主妇要去几趟洗衣店，用什么肥皂的问题，当然也有在城市当中疾病如何传播的重大问题。此项研究，伊格尔和他的同事们利用非洲预付费用户的位置信息和账户的余额来结合分析，发掘出成正比的是资费和收入，预付费越多的人往往都是富有的人。除此以外，他们还有一个相悖于直觉判断的结果，事实上贫民窟除了是贫困中心外，还会成为经济繁荣的跳板，这当中最重要的一点是，他是在间接利用手机的位置信息，本身和移动通信之间没太大的关系，而这些数据最开始是为了移动通信而产生的。总而言之，一旦有了数据化的位置信息，所有信息的新价值都会应运而生。

数据化的沟通方式

另一个数据化的前沿就是个人化，这牵涉到了人类的关系、经历和情感。很多社交网络公司的主心骨就是数据化的构思。人们在社交平台上寻找和维持朋友和同事关系，每个人日常生活中的无形元素都被提取出来了，从而有了新用途的价值。就是因为这个，有了将关系数据化的Facebook，从前被视为信息存在的社交关系，直到Facebook“社交图谱”的出现才开始被界定为数据了。通过创新的Twitter也创造了一个让人们容易记录和分享自身想法的平台，这在从前总是会成为人们遗忘在时光中的碎片，也因此让情绪有了数据化的可能。过去漫长的经历通过

LinkedIn而经历了数据化处理，好比是莫里转化旧的航海日志一般的过程，信息化为现在和未来的预测提供的依据，不论人们认识的是什么人，又可以在哪里找到一份心仪的工作等等。

可惜使用数据的技术还没有成熟。拿Facebook为例，用户得知自己的数据过早被泄露而产生了过激的反应，不过此时的Facebook精明地选择了忍耐。除此之外公司还在继续着数据收集的工作，也涵盖了商业模式和政策上对隐私问题的调整。目前，它面临的指责主要是来自采集了什么，而不是它利用这些数据去做了什么。

它的潜在用途非比寻常是毋庸置疑的。部分消费者信贷领域的创业公司正在打算将自己的信用评分依据转为Facebook的社交图谱。FICO的信用评分系统就正在打算通过15个变量来分析一个贷款者是否有偿还债务的能力。一家匿名的高风险投资公司有一项内部研究，结果显示呈现正相关的是个人偿还债务可能性和他的朋友偿还债务的可能性。老话说得好：物以类聚，人以群分，所以说Facebook正在朝着FICO的方向发展。很明显，在社交平台上的大量数据已经成了放飞想象的新型商务基础，它的意义早已超越了人们的照片分享、状态上传还有“喜欢”按钮。

Twitter也在同时间开始挖掘数据的新用途。从某种意义上来说，2012年每天超过1.4亿的用户发送的4亿条微博就好比是口头的随意零碎。它们通常的情况就是这样，可尽管是这样，Twitter公司还是实现了个人表述数据的数据化，这是从前从未有过的。和其他两家公司Data Sift和Gnip一起，Twitter公司达成了数据出售访问权限的协议。很多公

司对人们所发的微博进行句法分析，还有另一种情感分析的技术，这些技术用以获得反馈意见的汇总或是营销活动有着非常有利的判断。

位于英国伦敦的两家对冲基金Derwent Capital和加利福尼亚的MarketPsych也着手对微博的数据文本进行分析，以此为股票投资的信号（此前这项商业秘诀他们从未公开过，不知道是为了倾向良好公司的投资还是做空）。此时两家公司都开始向经营者出售信息。对MarketPsych来说，和Thomson Reuters合作在119个不同的国家提供了18864项的独立指数，像是每分钟更新的心情状态，有乐观、忧郁、快乐、害怕等等，还有很多像是创新、讼诉和冲突的情况出现。

被人类利用的数据频率远低于计算机，很多在华尔街以“金融工程师”知名的数学奇才，他们把数据送到了自己的算法模式中，以此来找到有效利用且能盈利的隐性关系。“社交网络分析之父”贝尔纳多·哈柏曼（Bemardo Huberman）的分析说明，在微博中出现单一主题的频率对于预测不少事情有着重要的作用，例如好莱坞的票房。他和一个在惠普实验室工作的同事一同开发了一个程序，勇于监听新微博的发布频率，并在此基础上他们来推测一部电影的票房，据说这方法要比传统的评估准确得多。

类似这样的数据用途举不胜举。Twitter微博的发布字数被限制在了140字，可是与之相联系的元数据却是非常丰富的。Twitter的元数据指的是“关于信息的信息”当中包含了33个分离项。其中有一部分的信息用处不是很大，好比是Twitter用户所用的墙纸，还有很多用户用来访问这

项服务的软件等等，可是剩下的元数据价值就很大了，例如他们在服务当中用的语言，所处的地理位置，还有关注的人和粉丝的名字、数量等等。《科学》杂志在2011年刊出了一项研究表明，世界上不同国家、不同地方的人们每天、每周都在用类似的模式。这结果是来自对84个国家的240万人所发布的5.09亿条微博的分析，从前这样的研究是无法想象的。可见，情绪也开始被数据化了。

态度和情绪因为数据化而成了可分析的形式，还可能转化为人类的行为。只不过要跟踪这些行为很困难，尤其是在大的社区和它的子人群中。

早期的发现可以用来推测数据化的发展方向。同谷歌一样，不少社交网络平台也拥有大量的数据。只要能深入分析这一类数据信息的话，无论是哪家公司都可以很容易地获得各行各业的动态信息。

企业竞争力的关键——大数据

越来越多拥有强大竞争力的公司都尝到了大数据的甜头，行业结构因此发生了翻天覆地的变化。只不过每一个公司的变化不尽相同。最有可能成为赢家的是大公司和小公司，而中等规模的公司则很难在这一次竞争中赢得胜利。

尽管在行业中亚马逊和谷歌还始终占据了第一把交椅，可是有别于传统的工业时代，它们并不因为有庞大的生产规模而获得企业竞争力。尽管技术设备还非常重要，但是核心竞争力也不是它们。公司要根据实际发展来逐步调整计算机的技术力量，逐渐转变固定投入为可变投入，

大公司技术配备的规模优势也因此被削弱。

小数据时代的赢家，包括那些线下的大公司，诸如沃尔玛、联邦快递、宝洁公司、雀巢公司、波音公司等等，都接到了大数据提出的挑战，它们纷纷认识到了大数据的价值和能力，并开始有效地收集数据和使用数据。与此同时，在科技创业行业和新兴行业中的老牌企业也开始了自己的大数据时代。

过去的十年间，劳斯莱斯作为航空发动机的制造商在产品分析的使用过程当中收集到了不少数据，从而完成了商业模式的转型。就在英格兰德比郡的劳斯莱斯运营中心长时间地把控着全世界超过3700架飞机的引擎运行的监控工作，它们的工作就是在故障产生之前找到问题。劳斯莱斯将原有的简单制造转化为附加值很高的商业行为依靠的就是数据。劳斯莱斯本身出售发动机，与此同时还用计费的形式为航空公司提供监控服务。现在世界上的民用航空发动机有70%的收入都来自提供服务所赚得的服务费用。

小公司在大数据时代也有了自己的春天。埃里克教授曾说过，灵活聪明的小公司会获得非固有资产规模的恩惠。换句话说，小公司因为没有过多的固有资产因此不具备很强的存在感，也因此在传播它们的创新成果上成本很低。最重要的是，由于大数据的最佳服务都基于创新思维，因此对于它们来说不一定要有大量的原始资本投入。数据不会被占有，但是可以授权，在云处理平台上快速且低成本运营的数据分析，在数据带来的利益当中，授权费则不过是抽取的一小部分而已。

大数据能让大大小小的公司都从中获利，它不仅仅是适用于使用数据的公式，对于掌握数据的公司而言也是适用的。拥有大数据的人总在尽可能地用各种方式加大数据的存储量，以换来最小成本基础上的利润。第一，它们的存储和处理数据基础已经很好了，第二，融合数据库的价值是独特的。第三，假设只要从一个人手中获得数据的话，数据使用者显然更省心省力。实际情况却没有这么简单，另外还有一群在另一方的数据拥有者也因此诞生。数据价值的不断呈现，很多人都会以拥有数据的人的身份来展现自己的能力，很多他们手中的数据是和自己相关的，如购物习惯、观影习惯，包括医疗数据在内。

消费者从中获得的权利也越来越大。他们可以自主决定要把多少数据授权给哪些公司。只不过不是所有的人都认为要用高价转卖自己的数据，更多的人愿意做的是免费用数据来换优质的服务，像是亚马逊的优质图书推荐。对于大多数对数据敏感的消费者来说，个人信息的营销和出售仿佛同博客、微博发表一样自然简单。

只是所有这一切消费者所期望的绝非消费者的意识或是喜好的转变就可以催生的。如今消费者将个人授权给公司，或者是公司从个人手中购买到信息，这程序都过于昂贵和复杂。无疑一部分中间商出现了，他们先是从消费者手中购买信息，再卖给大公司。只要有低成本消费者就非常信任中间商，那么就会产生个人数据市场，在很大程度上个人就成功转型为数据拥有者。美国麻省理工学院媒体实验室个人数据分析家桑迪·彭特兰就和一些朋友创办了ID3公司，他们的工作就在积极地将这一

切变为现实。

这些中间商开始运营起自己的工作了，只有当这些数据真正被使用时，消费者才算是真正意义上的数据拥有者。现在消费者积极地在呼唤着能够有足够配备的中间商的出现，而在此之前，他们不希望自己的信息被披露得太多。总而言之，只要条件成熟，消费者就可以从完整意义上转型为数据掌握者了。

中等规模的公司从大数据中获益并不大。波士顿咨询集团资深技术和商业顾问菲利浦·埃文斯（Philip Evalls）提到，规模优势最大的是超大型公司，小公司则具备灵活性。传统行业当中，大公司的灵活性不如中等规模的公司，而小公司的规模不如中等规模的公司。当大数据时代到来，公司的规模不要求太大，甚至不用足够支撑它的设备投入。往往是那些灵活的小公司才会在大数据时代中获得更大的成功。

国家竞争力也会被大数据所撼动。发展中国家越来越多地开始承接制造力，其他国家更在意行业创业的时候，工业化国家对于大数据技术的重视在逐渐提高，它们也因此始终在全球竞争中占优。不幸的是这优势只是暂时的。好比互联网和计算机技术，世界上绝大多数的国家都开始应用这些技术的时候，西方大国的技术优势就在慢慢消失。发达国家的大公司它们所获得的利好消息就是大数据会加剧优胜劣汰。凡是掌握了大数据的大公司，就可能在短时间内抛下自己的对手，遥遥领先。

尽管有这么多的好处，还是有不少隐忧的。大数据对全球各类事情和个人所处的位置预测得越准确，越是会侵犯到个人的隐私，似乎很多

人都还没有做好相关的准备。在一个数据充裕的阶段到来的时候，很多认知和制度都还没有适应，毕竟它们都来自小数据时代。下一章节就来谈谈大数据的弊端和不足。

第八章 让数据张口说话，管理决策中的大数据

管理不再是管理专家的专利，因为大数据时代的推动，数据取代了传统直觉式的决策管理，数据开始占领话语权的阵地。当企业管理转化为数据之后，客观数据统计结果才是最终成就企业，规避风险的利器。

客观数据最具发言权

二战时期的美国曾经在执行飞行任务上动用了大量的轰炸机，不少飞机都被地面的炮火击落，造成了大量的损失。飞行员的经验判断是由于主机翼的面积太大，地面炮火容易击中，可是如果在那上面再加装装甲，势必会导致重量过大影响飞机起飞。那要如何能既保护飞机，又能装上少量的装甲呢？为此美国军方内部的争论始终没有停止，最后只得请到了著名的统计学家沃德教授。

沃德先给每一名飞行员都发了飞机平面图，并提出让飞行员在平面图上画出弹孔的位置。飞行员画完以后上交时，沃德发现大多数的弹孔都集中在一张图上。那里几乎被打得千疮百孔，尤其是前机翼，可是在驾驶舱和飞机的水平尾翼上的弹孔却少很多。

看到这里，空军人员的意见就更为坚定了，他们要求在前机翼加上装甲，可是沃德教授的意见与之相左，他要求在弹孔最少的驾驶舱和水平尾翼装装甲。有一名飞行员出身的将军对此很不理解，说道：“教授的统计说明这两个地方最少受到攻击啊。”沃德教授回答道：“我们做统计的飞机都是安全回来的。从统计的结果来看即便是前翼被击中多次还是不影响飞行的。而这次统计很少被攻击的部分绝非没有受到攻击，反倒是因为被击中之后就坠毁了，无法统计。”将军听完仍然不同意，继续说：“尽管您的推断很有道理，但是从我的经验来说，当我还是飞行员的时候，机翼曾多次中弹，面临坠毁的危险，我的经验挽救了我自己。显然中弹概率最高的是机翼，我还是坚持我的观点。”

空军司令见两人的意见不统一很是为难。最后他还是决定接受沃德教授的意见，改进现有飞机的构造，在驾驶舱和水平尾翼后加装装甲。在采取了这一系列措施之后，盟军飞机的生存率大幅提高。为了确定这一策略是否正确，军方特地派出了特种部队和敌后抵抗组织到德国和法国等盟军飞机坠毁的现场去检测。结果显示不少飞机的坠毁确实是由于驾驶室和水平尾翼中弹导致的。

这次意见的对决很显然是一场科学方法和经验主义的对决，最后科学家的科学方式解决了问题。社会不断发展，大数据时代已经来临，很多数量化和科学方法从前无法解决的问题如今都有了解决的方案。大数据通常是借助计算机、各种数据采集工具、服务器、ERP信息系统等IT技术来获得大量的数据，再借由各种计算公式、统计方法找到解决的办法。

先来介绍一个数据解决疑难问题的例子。风险投资公司的名声不好，这往往是依赖本能感觉或是经验来做出的判断。风险投资的业务总是和各行各业相关，若是想要有一个量化的计算公式不太容易，更别说对各领域投资项目进行分析了。投资毕竟需要十分谨慎的态度，投资者寻求的方法必须是科学的，要去给有潜力的创业者和创业公司进行分析，评级和打分。

创业者的思想在投资者看来，要依据来寻求支持的创业者手机上的应用程序，浏览过的网站，还有读过的书籍来分析。首先了解他的大脑装了哪些内容，创业者阅读了多少书籍，在社交平台上又发布过什么样

的内容等等。在投资者看来，创业者在社交平台上所发布的内容如果是有知识有水平的话，譬如引用了多数的名篇名句，创业成功的指数就会随之增加。

著名的风险投资家克里斯·萨卡曾说过，分析投资项目的时候，他会先去看看创业者在Twitter上最近发的50条信息，以此来推测创业者是什么样风格的人。他应用了精心设计的投资算法，创业者的表现被纳入了计算公式当中，数据的分析让他看到创业者是否值得信赖。

投资者还会依照创业者手机号码是何时开始使用，以及每天的第一通电话拨打时间来分析和推断。例如手机的使用年份可以揭示创业者的人脉资源和人际关系，年龄大约是多少，是否有亲情号码计划或是座机转为手机的经历。手机号码的使用年份一系列信息的处理，投资者能看到更多关于创业者的信息，整个过程是合法的且完全保密地进行。最后投资者理想中的创业者应该具备年轻的、稳定的，有中产阶级家庭背景的，且创业有家人和朋友支持的人。

挖掘潜力股的数据化进程

最重视数据的NBA，当中有众多运用数据的高手。

大多数在NBA征战的球队都有专属自己的数据统计人员，不过每支球队的统计团队规模不一，这和球队对数据的关注程度高低有关。其中最为重视数据的两支球队非凯尔特人和火箭莫属。此外在联盟排在前列的球队一般都有数据分析团队，那些不重视数据分析的球队总是在联盟中垫底。美国知名的体育电视台ESPN篮球比赛分析师翰·霍林格曾说

过："NBA在这方面的领军人物必然是火箭的总经理达雷尔·莫雷。"

2006年，才满32岁的达雷尔·莫雷受聘成为休斯顿火箭队的总经理助理，第二年他就正式出任总经理，当时为人们称为"少帅"。达雷尔·莫雷毕业于麻省理工学院的MBA，身高超过了1.95米的他居然没有任何的职业篮球经历。莫雷曾说过："大学时期的我，任何一项运动都勾不起我对数学一般的兴趣。"

正因如此，莫雷本着对数学和统计分析的偏好，在获得麻省理工学院的MBA学位之后，创立了属于自己的体育统计和分析的数据研究公司。从那时起，莫雷就和篮球相关数据结下了不解之缘。凯尔特人特别聘请他管理球队的球票销售。工作十分出色的他在2002年就被凯尔特人队聘请为篮球管理运作副总裁。莫雷在这个职位上用数据统计和分析的方式选中了当时名声不那么响亮的韦斯特、杰弗森和杰拉德。一时间争议很多，但是此后三人的表现说明莫雷的选择是正确的。

加盟火箭队之后的莫雷更是引入了一种同步数据录入的技术，这样NBA最新最全的统计就可以获得了。在传统数据分析模型的基础上，莫雷带着自己的团队还会使用比赛录像模拟技术，简单说就是电子游戏的方式来演练。而这一切都是计算公式所编排出来的规则演练程序。就好比当火箭和国王在就前锋阿泰斯特交易谈判过程中，莫雷和他的团队用EA Sports的NBA游戏整个模拟了阿泰斯特加入火箭队之后的比赛情况，这才最后做了决定。

除此之外，莫雷还会通过统计学的模型对每一个上场球员的综合表

现进行追踪和评价。莫雷对于自己是如何分析数据的从来都讳莫如深，可是所有人都明白他手中有一套完整的计算方法，其中最厉害的莫过于预测一个球员同球队配合时的表现。莫雷说过："一个球员和球队的合作会是什么样的，我们无法决定。例如在棒球比赛当中，中场手只会是中场手，而在篮球比赛当中，一个球员在球队中的表现无论如何要想办法来预测。"

莫雷接手了火箭队以后也曾遭遇了很大的挑战，队中姚明和麦蒂的薪酬就占去了整个球队的40%，两人的状态也不是太好，因此球队在购买其他大牌球员上总显得有些捉襟见肘。可是莫雷想出了一个办法，他用自己独到的数据分析方式挖掘球队潜在的优秀球员。

2007年，莫雷在选秀当中放弃了当时最红的杰伦·戴维斯和约什·麦克罗伯茨，转而选择了身高只有1.80米的后卫亚隆·布鲁克斯。很多人对莫雷的举动表示不解，可是到了夏季联赛布鲁克斯的表现却让大家难以置信。最后布鲁克斯在那一赛季的表现压倒了当时包括"榜眼"凯文·杜兰特在内的诸多新秀，冉冉升起成了一名篮球新星。

莫雷后来又利用自己的方式引进了巴蒂尔、斯科拉、阿泰斯特等"价廉物美"的球员，火箭队在莫雷的带领下即便是麦蒂和姚明轮番缺阵的2008年1-3月，还是取得了22连胜，闯进了季后赛。

时代因大数据而变革

统计学才是数据化的技术保障。统计学本身是一门基于搜索、整理、分析数据等技术手段，来推测对象本质或是预测对象未来的综合性

科学。也就是寻找挖掘的对象，用数字标志来记录对象的现象，再汇总数据找到事物的本质和规律。统计学最为核心的思想是要汇总偶发性的一切事件，再找到必然的规律。很显然，一个事件的发生常常是偶然的，但是大多数事件叠在一起就会有必然性产生。

企业管理不论是从市场运营出发还是对人的管理出发，只要认真地审视每一个客户或是员工的表现，叠加在一起，规律就会自然而然呈现出来的。

统计学的发展经历了三个时期。最早统计学的概念产生于古希腊的亚里士多德时期，早期是用于计算和统计各个城邦王国经济发展水平和居民生活情况的。如今，在国家管理、企业运营和科学研究等多个领域统计学都在被广泛应用。在整个社会当中，统计学的发展大致可以分为以下三个阶段。

第一个阶段是城邦政情。有观点认为，最早应用统计学方法的记录是古希腊亚里士多德所撰写的“城邦政情”。早在公元前8世纪，古希腊就建立起了不少城邦为基础的国家。亚里士多德主张科学和数学管理，在抵御波斯帝国入侵的问题上，它们对各个希腊城邦的人口、军队、经济和粮食等方面进行统计，用于分析抵御外敌入侵的能力。很快这方法就传遍了欧亚各地，到了17世纪就发展为真正意义上的统计学了。

第二阶段是政治算术阶段。这一阶段最大的特点就是把统计方法和数学推算、推理相结合，对社会经济问题进行分析，其中更为注重定量分析方法的应用。直到现在，早期的“城邦政情”已经发展到了对国家和

社会所有经济、社会发展的统计分析，现在大家所熟悉的GDP、CPI、基尼指数等等都是统计学上的重要数据指标。

拿基尼指数来举例。1922年意大利经济学家基尼提出了基尼指数，用于在劳伦茨曲线基础上来判断收入分配的公平程度。比例数值介于0和1，越是接近0就说明收入的分配越公平，反之则越不公平。长期的数据统计结果表明，只要一个国家的基尼系数超过0.4，就说明这个国家贫富差距大，社会不稳定因素多，常常会有失业、罢工等危机事件发生，如果高达0.6的话，那就说明收入相当悬殊。

第三阶段是大数据时代。在计算机和现代检测技术发展的今天，数据的提取已经不是难事。通过计算机技术，一个银行每天的客流量和业务办结情况可以随时进行统计。与此同时多个银行的信息还可以汇总到一起来进行分析。有了软件技术，客户的身份信息也可以快速转换为年龄、地址等等数据，很多的客户和业务都能被贴上信息标签。具备了这些，企业每天都可以累积庞大的数据，然后再通过网络和检索汇总成海量数据。因此大数据的概念产生了。

那些数量已经超越人工或是简单的计算机软件处理能力的庞大海量数据就被称作是大数据。例如有商场客户购买情况，搜索引擎上网民对某一条词条的搜索情况记录，还有全国性的人口普查等等。海量信息的出现，管理者不得不放弃原有单纯分析销售额和客户成交率的做法，转而去针对不同年龄、不同性别和不同地区的客户成交率的计算。

关于温度、压力、湿度、亮度、粉尘颗粒等信息检测技术会越来越

方便，海量的数据就此被积累起来，用来对人们在不同条件下的行为变化进行分析。例如中国互联网络信息中心曾有过这样一项统计，到2012年12月底，中国的网民数量已经达到5.64亿人，普及率高达42.1%，和2011年底相比提高了3.8%。中国的手机网民的数量大致为4.2亿人，和上一年度相比增加了大约6440万人。2012年70.6%网民用台式电脑上网，和上一年底相比下降了将近3个百分点，相比之下手机上网增长数量比较快速，由上一年度的69.3%提高到了74.5%。此外百度的统计也显示，2012年度网民不同时间的上网活跃度依次为：20：00—21：00（6.53%）、21：00—22：00（6.44%）、16：00－17：00（6.20%）、19：00—20：00（6.13%）、15：00—16：00（6.09%），这当中在晚饭过后的3个小时（19：00—22：00）是网民最为活跃的上网时间段。

大数据现在还促成了另一种概念，即大量数据的运用统计，也就是从看似不相关的事物中去寻找有意义的规律。最近有一则报道，称一名专业人士特地去上海均价4万元以上的高档小区做物业管理，每天都拿着扫描枪对着小区的每一件带条形码的垃圾进行扫描，晚上再一一整理出来，以此来了解这个小区的业主喜欢喝什么水，吃哪种油，穿什么牌子的衣服。一段时间的数据积累后，就充分地了解了该小区业主的消费偏好，形成系统之后再将报告转卖给大公司。

京东商城的老板刘强东在看过统计报告后发现，如果女性用户大于50%的话，就可能盈利。其实这当中的秘密是，男人和女人的购物偏好

不同，男人买的都是毛利润“低”的产品，女人则是购买毛利润高的服装、鞋帽和化妆品等等。他说：“我们的目的是要做一个大型综合购物平台。如果只有男人光临的话，那这个平台是有缺陷的。”

大数据时代所带来的变革是和生活、工作、思维都密切相关的。企业管理者必须具备一双能够穿透数据看到本质的眼睛。商业模式在数据时代所受到的冲击会直接对企业的经营理念、策略制定以及衡量标准有非常大的影响。只是企业经营的核心是永恒不变的，永远都是洞察客户的需求，为他们提供最合适的解决方案，这样才能创造价值。大数据时代让每一个人离完美目标的实现更进一步。大数据思维的营销实际上就是“向数据中求答案”。

大数据时代的风险规避策略

第一是要通过评估来挖掘潜在的风险。现代企业十分关注产销互动，也就是说在供应采购、生产单位和市场营销三者间要彼此通报信息。也就是这样部门和部门之间才能协调好规划。

对采购物流部门而言，即便是在供应商供货正常的时候也要居安思危，考虑一旦断货或是交货不及的话该如何。惠普公司的营销团队每天都会收到黄色警报和红色警报的邮件，它们均来自生产部门。所谓黄色警报就是在警示最近某一类产品的配件存在缺货的可能，产品的发货可能会延迟，这样的话销售团队就必须在与客户签订合同时进行协商，或者建议客户选择其他的产品。而红色警报则说明某种产品的存货相当紧张，濒临断货，销售团队对客户要停止报价，避免签订了合同却无法供

货。还有采购、物流部门为了尽可能地了解未来的走势，一定要事先了解供应商的动态信息，包括供应商所在领域的经济态势，以免在断货的时候出现供不应求的局面。惠普在管理供应商时所采用的原则是“少数及密切配合供应商”。无论是商品管理、质量和工艺管理等等，都要培训供应商，以改变其内部的流程。戴尔还和自己的供应商分享品质管理的工具，以提升其自身采购的管理水平。

惠普用一套量化的评分体系来评价供应商，每个季度都会评估供应商，获得结果用来分析将来每个供应商的表现，事先做好准备。

每季度对供应商所做的多角度评价，让惠普能对所有供应商做一些指标评分上的排序，找出最优的供应商，这样一来，未来可能存在的风险就被规避了。当然除去评估的排序外，如果惠普发现供应商某一项得分过低，就会即刻淘汰。所以惠普的供应商在每一个环节都做得非常好。

第二是对销售市场最为科学的预测。曾经某一家代理商在一段时间内业绩增长非常快，而另一家代理商则始终没有增长。还有好多这样的例子，诸如一个年轻的销售员成了一个大企业的负责人，销售业绩大幅增长，反倒是一个负责小企业的资深销售员却无法提升企业的业绩。

看到这样的例子，就该明白必须去衡量一下销售员的工作到没到位，销售员的能力是否合格呢？这当中要使用到的是一个称作“销售预测”的重要工具。所谓预测和销售人员了解自己的业务有关系，从而判断自己在客户中的接受程度高低，找到和目标之间的差距，预先采取行

动。

评价销售能力最重要的方法就是销售预测。市场销售人员的预测可以帮助他们更接近客户，更为扎实地完成市场工作。此外，企业运营使用销售预测也是十分有必要的。

预测市场的时候别忘了一个很有用的工具，那就是销售漏斗。它其实是一个系统记录客户业务机会的工具，预测未来业务机会的多寡可以用它，记录不同业务部门和不同时间的业务机会也可以使用它，包括营销团队的风格分析也少不了它。说到销售漏斗主要有以下几个特点：

1.销售漏斗可以记录业务机会，当中有客户的联系信息，譬如联系人、电话、传真和邮箱等等，还有业务机会的预期成交时间、成交金额和产品品类等等。

2.销售漏斗中还有“成交概率”，它是销售漏斗中的核心所在。成交概率一般可以分成以下几种比例：

·0%，彻底没有机会了；

·10%，比较了解客户信息；

·30%，上门拜访客户并赢得认同；

·50%，客户主动表示认同；

·70%，客户主动做出承诺，投标中占据了优势地位；

·90%，完成签约；

·100%，客户已完成付款。

3.未来业绩通过加权计算完成。每个业务机会都能把预期成交金额

和成交概率相乘，结果就是未来的加权预期金额。

预期成交金额×成交概率=加权后成交金额

4.一系列的业务机会都可以通过加权预期金额开始统计，未来一段时间内的业务就可以由此进行整体判断。加权预期金额如果比目标还大，那就表示业务已处于正常情况，反之则需要适当调整和改进。

5.若是每个团队一个月实际完成的是加权预期金额的七成的话，那证明团队预估的结果过于乐观。他们若是预期下一个月是900万元，那么最终很可能只能完成630万元。

企业文化的数据化构建

先来说说由主观向客观转变的圆桌会议。西门子在评价人的时候所用的方法很是独特——PMP（performance management process）。如果将PMP划分的话可以分成“PMP圆桌会议”和“PMP员工对话”。说白了，PMP就是每年都把人力资源部和业务经理聚在一起，细细讨论每一位员工的未来发展，再对其后续的晋升和发展做出决定。

每年一次的圆桌会议，与会人员是公司的中高层经理和人力资源顾问。在圆桌会议上，业务经理都会邀请自己上一级的经理和其他业务经理评估自己的团队，在综合所有人的意见后，再和人力资源部一起给员工打分。第一步是要人力资源部把所有员工的名单拿出来，挑出一个来，把所有和他有关的绩效列出来；第二步则是直属经理阐述该员工平时的表现，从九项能力对员工进行点评，有了解这名员工的也可以说出自己的观点。如果没有大的分歧的话，直属经理对它的评价就将成为其

评分的主要依据。如果有大的分歧的话，就会相应地进行协调。尽管人人都是主观的感受，但是综合多个人的意见之后主观意见就会变得客观。此外西门子在评价员工时，还会把评价的结论落到实处。

“圆桌会议”的内容主要有：

·评估员工发展潜力。

·阐述员工上一年的业绩。

·如何调整员工薪酬问题。

·针对员工制定具体的发展策略，例如调整岗位，授权更大的职位，参加培训，指派教练，还有轮岗等等。

·发展措施。

圆桌会议之后，员工就会一一和经理谈话，他们会获得自己的评估结论，经理会告诉他们评估的结果是具备潜力，那么将来有一天他们可以获得晋升。员工得知了以后必然获得鼓舞，有了相应的自信，并相信自己在西门子会有美好的将来。有了这些自信后就会主动地去承担更多的工作和责任，获得更佳的业绩。在西门子看来，员工潜力的评估和认同，包括和员工之间的交流都可以激励员工，并增强员工的忠诚度，这是一种非常有效的发展措施。

再来说说定制式的培训体系。西门子遍布全球的西门子管理学院成了为员工提供及时管理培训的重要资源，员工经过培训管理能力能有效提高。围绕西门子的战略和文化，西门子的培训系统在西门子的能力模型基础之上建立。西门子每每遇到战略变革的时候，西门子管理学院的

作用就相当重要了。它会依照形势的变化而制定能力模型培训课程。上述的九项能力都有相对应的课程，在西门子案例和风格的结合下进行讲解。

通过“变革管理”的培训，管理学院通过西门子的真实案例来进行教学，学员往往会扮成变革中的人员角色，带着自己的背景材料。随后学员依照自己所扮演的角色开始讨论和演练，对各方如何协调进行分析，最终达成目标。西门子的培训师梯队的组成主要是内部的培训师，这样做的优势在于他们十分了解西门子的企业文化，完全可以利用西门子的思维来讲解培训内容。

培训当中，西门子很是强调量化培训结果的评估。评价培训本身是一个非常主观的感受，西门子利用合理的评估方式来化主观为客观。第一，直接参加训练的人员，在培训两周之后，西门子管理学院要给学员们发邮件，学员开始填写问卷，然后统计分数。这样做的目的在于避免培训才结束后的情绪化的影响，学员可以在培训后理性思考自己获得了什么。除了这些以外，管理学院的人还会调查参训人员，询问在学习之后，学员们有哪些改变。分数不高于85分的时候管理学院就必须重新反省课程，数据会显示问题的关键是老师还是课程本身。

说说企业文化建设的数据检验问题。每一年西门子公司都会全面进行员工满意度的调查，并从中收集员工对企业战略实施的意见，是否了解和执行企业文化，甚至对管理水平的评价等等。量化评价员工的调查结果是大多数企业使用的科学方法。

西门子为了确保员工调查结果的客观性，特地请来独立的调查公司对员工做调查，西门子还和调查公司签署协议，员工所填写的问卷即便是西门子的最高层管理者也不被允许去看。因为西门子知道只有这样才能让员工说真话，不因为担心被打击报复而说实话。通常调查会有30多个问题，有公司战略、文化和管理等多方面的调查，每个问题都要用0-10分来打分，分越低表示越不满意，越高说明越满意。

例如在文化调查中的问题就有：

（1）“你是不是愿意举荐他人来这个公司呢？”之所以问这个问题，是想了解员工对公司的认同度，举荐他人来公司说明他热爱这个公司的环境。

（2）“假如有其他公司的薪酬、职位和权限和现在的公司一样的话，你还愿意继续留在公司里吗？”问这个问题的目的在于，公司当中难免会有很多员工认定公司的依据是钱和职位，并不关心公司的文化和环境，一旦有工资待遇和现在差不多的公司出现的话，还能继续留在公司中的员工才是真正认同公司文化的。

（3）“公司的文化价值理念是否体现在同事的行动上？”个人表现不是企业文化，企业文化应当是集体的表现。所以要知道员工的行为是不是公司文化的体现，就必须了解广大的员工是不是已经了解了公司的文化价值。

管理总是紧密地和文化相连的，而衡量员工是否认同管理环境也是文化的职能，它也说明了管理者是不是能依据西门子九项能力来工作。

这其中存在着一个核心问题："你是不是愿意举荐他人成为你上司的下属呢？""你的上司是不是常常有创新的思路呢？"

那么调查员工企业文化的工具是什么呢？西门子和众多大公司一样，习惯在调查员工当中应用一个标杆模型，那就是盖洛普的Q12。

著名的国际咨询公司盖洛普公司曾经对12个来自不同行业的24家公司的2500个经营部门的数据进行收集，为的是发现企业文化的建设，随后又针对10500名来自不同公司和企业文化的员工态度统计分析，结果表明员工保留、利润、效率和顾客满意度的指标中存在12个关键问题，这些均是公司文化氛围和管理水平的反映，这其中有12个经典的问题，如下：

1.我的工作我是否知道有要求？

2.做好工作所必须具备的设备和工具我是否都具备了？

3.每天我是不是都有机会在工作中做自己擅长的事情？

4.我在过去的七天中是否因为出色工作而受到表扬呢？

5.我的主管和同事是不是关心过我的个人情况呢？

6.我的发展是不是有同事鼓励我呢？

7.我在工作中的意见是不是经常受到重视呢？

8.我的工作是否由于公司的使命目标而重要呢？

9.我和同事们是不是都在为高质量的工作而努力呢？

10.工作中我是否有一位最为要好的朋友呢？

11.过去的半年里，是否有同事找我谈过我的个人进步呢？

12.过去的1年中，工作中的我有没有机会去学习和成长呢？

上面12个经典的问题，在具体企业应用中根据实际情况进行适当的调整，并转换成员工的调查，企业的现实情况就因此能显现出来。调查完员工之后，再依据结果显示出来的数据，对企业的状态进行分析，还利用同比、环比、邻比等方式开始比较，再以此找到缺陷和不足，决定要采取的行动。

假设有一家公司的员工满意度调查，其中关于“你愿不愿意举荐他人到自己的公司里工作”这一题的数据显示，2012年中国分公司的员工数据为83分，而总部的这一数据则是92分，两者相差很大，但是同相关行业水平78分进行对比的话，就会发现该公司的工作环境还是比较好的。

再来说说“过去的半年里，是否有同事找我谈过个人进步”这个问题的结果，其中数据表明，中国分公司经理对员工的关注太少，也说明下一步必须对此进行重视。

第九章 更自由，更开放，大数据的机遇和挑战

大数据为所有人都提供了更为开放、自由的发展机会，但与之相应的，一起到来的还有人类从未碰到过的挑战。一切传统规则被大数据时代所颠覆、推翻和打破，又该如何在保证正确的发展方向的同时避免不必要的风险和负面影响呢？

人机结合的未来发展趋势

人机结合，或者说是人和数据的结合是未来的一个进步发展趋势，数据经过人类的改造会越来越智能。

不少人会有疑问，目前的大数据究竟发展到了什么阶段？显然这个阶段的水分还太多。大数据对社会方方面面的影响已经很大了，这点毋庸置疑，只不过它的爆发点还没有到来。大数据时代到来，数据开始和人的经验结合，两者彼此激活，人们的智慧也因此获得了规模化的放大，因此整个社会的发展都因为有了大数据而发生巨大改观。不过大数据应用未来的空间一定会比现在的成绩更为宏大，数据和人之间的“数据化运营”和“运营数据”的闭环系统也会渐趋完善，只不过在人机结合中的巨大空间还要进一步挖掘。

数据的种类在未来一定会超过人们的想象。从前更多的数据是集中在监控外部行为上，像是网络购买行为或是网上社交行为，皆为“远距离”的数据提供，但即便是这样，这些数据的运用还是不太好。随着可穿戴式设备的出现，人和数据就能真正意义上融合在一起了，就好像谷歌眼镜等等，我们所看到的一切都会瞬间被数据化，还有健康手环类的设备，包括能够深度收集脑电波数据的设备，人们未来的活动都会转变为数据。现在睡眠记录状况仍处在初步的应用阶段，可是不远的未来，人们每一秒的生活都会为数据所记录。

商业发展的更多机会是伴随着万物皆数据的时代一同到来的，那时候人们的判断可以依靠数据进行，例如什么时候是吃饭的最佳时候，什

么时候又是睡觉和记忆的最佳时候等等都可以利用数据来预判。即便是把人的各种数据都记录下来的话，每一个时刻适合做什么，要采取什么样的策略数据都能够预判出来。或许就在那个时候决定人们聪明与否的指标不再是IQ，而是是否拥有优质的数据。

未来的进步模式就是人和机器的结合，或是人和数据的结合，数据因此变得更加智能。

眼下最需要做的事情就是人机分工，人做人擅长的事，机器做机器擅长的事。

将来人机之间的界限越来越模糊，随着人机结合，越来越多的无人操作机器会不断出现，而在人类身上会有更多数据的血液流淌着。

数据时代，引发时代大变革

又一次因为大数据所带来的时代转型到来了。大数据和其他的新技术一样也必须经过硅谷臭名昭著的技术成熟度曲线，即在媒体和学术会议的大力宣传之下，新技术的趋势会在短时间堕入谷底，而让众多的数据创业公司出现危机。不过不论是过热期还是幻想破灭期，对于正确理解正在发生的转型重要性都是非常不利的。

感受宇宙可以通过望远镜，观测微生物可以通过显微镜，为了更好地理解世界可以通过收集和分析海量数据的技术，人们现在才意识到了这种理解世界的新方法。事实上，真正的革命并非去分析数据的机器，而是人们如何去利用数据和数据本身。

天文学的领域和其他各个领域一样都在发生着变化。人类在2003年

第一次破译了人体基因密码的时候，在工作了十年之后，三十亿对碱基对的排序最终完成。十年之后全球范围内的基因仪要完成同样的工作只需15分钟而已。金融领域也是如此，美国股市每天成交的股票多达70亿股，经由数学模型和算法基础的计算机程序自动完成的就有2/3左右，通常这些程序是在海量数据预测利益和降低风险的基础上进行的。

数据怕是要淹没所有的互联网公司了。每天谷歌公司处理的数据超过了24拍字节，也就是说所有被处理的数据量已经是美国国家图书馆馆藏纸质出版物的上千倍。像创办时间还不足10年的Facebook，一天照片的更新量都超过了1000万张，人们每天在网站上点击“喜欢”按钮或是书写评论的次数也超过了30亿次，很显然Facebook为了挖掘出用户的喜好用了大量的数据线索。谷歌的子公司YouTube每个月接待的访客数量也超过了8亿，几乎每一秒都有一段长约一个小时的视频上传。Twitter上的信息量也在每年翻番，到2012年，上面微博的发布总数已经超过了4亿条。

科学研究、医疗保险，还有银行业和互联网，几乎每个领域的故事都是相似的，爆炸式的数据增长所带来的故事几乎是重复出现。它的速度似乎已经超过了机器制造的速度，更是超过了我们的想象。

究竟我们周围有多少数据，增长的速度有多快？很多人都试图去找到这些答案。尽管不同的人用的是不同的测量方法，测量的是不同的对象，但结果都是大获成功。南加利福尼亚大学安嫩伯格通信学院的马丁·希尔伯特（Martin Hilbert）曾做过一个比较全面的研究，对象是统计出

所有人类创造、存储和传播的所有信息的确切数量，范围涵盖了书籍、图画、电子邮件、照片、音乐、视频（模拟和数字），此外还有电子游戏、电话、汽车导航和信件。马丁·希尔伯特还曾经以收视率和收听率为基础，开展了电视、电台等媒体的研究。

有意思的是2007年的时候，仅仅只有7%的存储是在报纸、书籍和图片上的模拟存储，剩下的都是数字数据。可就在不久以前，情况却发生了翻天覆地的变化。1960年，实际上“信息时代”和“数字村镇”的概念就已经存在了，可是到现在为止这些概念还都是新的。到了2000年，全球数据量中仅有1/4是数字存储信息，其他的都是在报纸、胶片、黑胶唱片和盒式磁带这些传统的存储介质之上。

最早期是没有太多数字信息的。对于经常上网和网上购物的人来说那是非常微小的一个部分。实际上，1986年，全球计算能力的40%几乎都用在了袖珍计算器上了，当时袖珍计算器的计算机能远高于所有的个人电脑处理能力。只不过随着数字数据快速地增长，很快形式就产生了翻天覆地的变化。希尔伯特曾经说过，每三年数字数据的数量就会翻一番，相比之下，模拟数据的数量则变化不大。

数字数据的增长确实是惊人的，或许信息存储的发展速度已经超过了经济增长的速度，从计算机数据处理的速度来看，已经是世界经济增长速度的9倍之多。这也难怪总有人在埋怨信息过量，而因此受到了极速发展的冲击了。

再往人类历史的前面看，如果拿1429年前后古登堡发明印刷机时的

信息存储来和现在的信息进行对比，在1453-1503年的50年间，历史学家伊丽莎白·爱森斯坦（Elizabeth Eisenstein）发现大约有800本书籍被印刷出版，这比1200年前君士坦丁堡建立之后欧洲所有的手抄本数量还要多。欧洲的信息存储量花了50年的时间在当时翻了一番，可是现在大约是3年时间就可以翻一番。

这能说明什么呢？谷歌的人工智能专家彼特·诺维格（Peter Norvig）曾在美国宇航局喷气推进实验室工作，他会把这种增长与图画进行类比。第一，他让大家去想象一下在法国拉斯科洞穴壁画上的标志性的马。从这些画的追溯来看可以一直往前追到一万七千年前的旧石器时代。然后再思考一张马的照片，或者是毕加索的画作，或许看起来和那些洞穴上的壁画差异不大。实际上，当毕加索看到那洞穴上的画作时就曾开玩笑说道："从这以后人类就没创作出什么东西来了。"

毕加索的话也对也不对。回想一下那壁画中的马，当时画出马需要很长的一段时间，可是现在已经不需要了。这就是变化，当然变化兴许触及的不是最核心的部分，因为他画的仍旧是一匹马。可是诺维格说过，只要去想象一下，现在的人们每秒钟可以播放的马的图片是24幅，这是一种源于量变的质变，也是一部电影和一幅静态画作的根本区别。大数据也是一种源于量变的质变。物理学家和生物学家的研究表明，规模被改变的时候，事物的状态也会发生巨大改变。

就拿纳米技术来说。纳米技术一直都在将事物变小，而非变大。这当中的原理在于事物有了分子级别之后，物理性质自然而然就会发生改

变。只要人们了解了这些新的性质，就能用一样的材料来创造此前做不到的事情。像是原本用来导电的铜，用纳米技术之后它就无法在磁场当中导电了。还有原本抗菌的银粒子，一旦以分子的形式存在的话，就会消失掉原有的性质。所以说在纳米级别上的金属会变得非常柔软，即便是陶土也会变得有弹性。事实上，人类所利用的数据量增加的话，很多在小数据量基础上无法完成的事情就都可以完成了。

人们有时会认为生活有着方方面面的约束，而这些对世间万物都存在约束力。可是真正的情况是，即便是相同的规律，人类感受到的约束力只是对人类有效而已。唯一对人类最为重要的物理定律无疑就是万有引力定律。这个定律每时每刻都在控制着人们，而细小的昆虫却不在乎这个。对它们来说有效的物理力是表面的张力，因为它们可以凭借这个力量在水面上自由地行走而不至于掉下去。事实上这个力却是人类不在意的。

生物体在不在乎万有引力和它的大小是有关系的。同理，信息的约束力也与规模有关系。之所以谷歌能够几近完美地给出一个和通过海量真实病例推断出的流感蔓延情况一致的结果，正是因为它能够比疾控中心更快地得到实时的数据。Farecast预测机票价格的走势，让消费者在经济上获利也是同样的道理。正是因为有了供其分析的数千亿的数据项才能让它们的服务功能如此强大。

虽然大数据时代才刚刚来到，但是每个人的日常生活似乎都已经离不开它了。例如能够自动过滤垃圾邮件的垃圾邮件过滤器，而即便它还

无法区分“发#票#销#售”和“发票销售”之间的区别；交友网站会重新搭配依照个人性格和之前配对成功情侣之间的关联；有“自动改正”功能的智能手机能通过分析人们此前输入的信息把个性化的新词收入到手机的词库里。不过这一切都只是个开始。这项技术很快就可以打败地球上的诸多东西，这一点从可以自动转弯和刹车的汽车，以及IBM沃森超级电脑在《危险边缘》游戏节目中打败人类就可见一斑。

数据可以表示世间万物，会带来惊喜

哪怕一点点的想象，数据化进程就会转化世间万物，还会带来意外的惊喜。IBM的“触感技术先导”专利和东京的越水重臣教授对臀部的研究工作理念是基本相通的。有知识产权的律师将其称之为是一块有灵敏触感的地板，犹如一块巨大的智能手机屏幕。它本身有着非常广泛的潜在用途，凡是放在上面的物品都可以辨认。最基本的用途就在于开灯和开门。除此以外更为重要的是，一个人的体重、站姿和走路方式也可以成为其判断身份的依据。某个人在摔倒以后是否站起来了也是它能知晓的。零售商在有了它之后可以了解商店的人流量。而数据化的地板能够有无穷无尽的用途。

事实上，这一切并不荒谬。一群健身迷和医学疯子，还有技术狂人发起了“自我量化”的运动，身体每一个部位和生活中的每一件小事都会因为数据测量而变得美好，还可以说是量化的方式使之更为美好。当前，这个运动的规模还不够大，但它一天天地在壮大自己的队伍。

智能手机和计算机技术的推广，使得个人生活行为的数据化已经变

得极为容易。很多创业公司为了研究睡眠模式测量了人们夜间的脑电波，例如Zeo公司就已经有了全球最大的睡眠活动数据库，用于解释男性和女性睡眠时眼睛快速动量的区别。还有Asthmapolis公司把感应器绑在了哮喘病人的呼吸器上，然后用GPS定位的方式来收集位置数据，再来推断病因中的环境因素。Fitbit和Jawbone公司则是对人们体力活动和睡眠进行测量。Basis公司利用腕带来检测人们的生命体征，像是心率和皮肤电传导率等等，这些都可以用来推断他们所承受的压力。2009年，苹果公司有了一项专利，也就是对血液氧合、心率和体温的数据采用音频耳塞的方式来收集。这个时代数据的收集似乎已经比任何时候都简单且不受约束了。

人们似乎从数据化那里得到了更多人体运作的方式信息。挪威耶维克大学的研究人员和Derawi Biometrics公司联合开发了一款可以用在智能手机上的应用程序，它主要的功能在于分析人走路时的步伐，并将其作为手机解锁的方式。与此同时佐治亚理工学院的罗伯特·德拉诺（Robert Delano）和布莱恩·派尔思（Brian Parise）也开发了一款名为iTrem的应用程序，主要可以用来监测人体的颤动情况，为帕金森症和其他神经疾病的病人服务。医生和病人都受益于这些应用程序，病人不需要再到医院去做昂贵的体检，医学家们更是可以远程检测病人的疾病情况和治疗情况。东京的调查人员提到，尽管智能手机上的这些应用程序和三轴测振仪相比还不那么精确，但是一般使用是没有问题的。这也再一次说明，完全精确并非总比不精确有效。

大多数的时候，信息会被采集存储后再来使用，任何领域，任何事情都可以用这样的方法处理。Green　Goose是一家创业公司，它们主要销售的是放在物品之上的微型运动感应器，以检测物品的使用次数。像是放在一捆牙线、一个洒水壶或是一盒猫食上，就能读出牙齿情结、植物护理和宠物喂养的所有信息。人们常常对“物联网”非常狂热，想在生活中的所有事物身上植入芯片、传感器和通信模块。好像这个词和“互联网”非常相近，但实际上它表示的是一种典型的数据化手段。

当下假设有一个重大的基础设施项目正在进行，很大程度上这和过去所做的一切都不一样，不管是罗马的水渠还是启蒙运动中的百科全书。它是全新的，但是又让人们置身其中，又好像是无形的，但和水渠中能触摸到的水不一样，因此人们总是无法意识到它的存在。所谓的它就是无处不在的数据化进程，它能够带来和其他基础设施一样的社会大变革。

城市的发展因为有了水渠而成为可能，启蒙运动因为有印刷机而得到推动，国家的兴起因为有了报纸而奠定了重要基础。所有流动的譬如水，譬如知识基础设施所依赖的。包括电话和互联网也不例外。相比而言，数据化所带来的转变则是革命性的。在大数据的帮助之下，世界在人们的眼里不再是一连串自然或是社会的现象了，而是由信息本质构成的。

一个多世纪以来，物理学家对人们宣称世界的本源不是原子而是信息。当然这个理论听起来有点让人难以置信。但是在数据化的推动下，

人们能够全面采集和计算有形和无形事物的存在，并对其进行处理。

世界被视为是信息，可以将其视为是数据的海洋，它为人们提供了一个前所未有的看待这个世界的角度，可以充分地渗透到生活各个领域的世界观。

早晚会有一天，水渠和报纸的影响会由于数据化而渐渐消退，与此同时，再为人类创造一个数据化时代下的产物。因此它对传统印刷术和互联网也产生了冲击。只不过目前它还只是在商业领域上被应用。

数据化带来的挑战前所未有

大数据的核心在于带来了分析信息的三大转变，它导致了人们理解和组建社会方法的改变。

第一个转变就是大数据时代的到来，让人们可以更多地分析数据，不再依赖于随机采样，而是分析和某现象相关的所有数据。19世纪以来，当有大量数据出现时，人们的处理方法往往都是采样，这种方法是在缺乏信息和信息流通受到限制的模拟数据条件下产生的。通常很多人会觉得这是理所当然的，可是当高性能的数字技术开始流行的时候，人们才明白这是人为的限制。和小数据范围相比，大数据的分析能够带来更高的精确性，而从中还能挖掘出不少此前无法发现的细节问题。所以说，人们利用大数据看到了更多样本本身无法揭示出来的细节信息。

第二个改变在于，大数据的研究和分析，让人们对于精确度的追求不再那么热衷。在测量事物能力受限的情况下，人们最为关注的事情莫过于精确的结果。要是购买者不知道牛群里的牛是80头还是100头的

话，那显然交易就很难进行了。一直到今天，人们还是习惯于在精确的基础上建立数字技术。例如只要电子数据表格能把数据排好，那么数据库的引擎就可以帮助人们检索出内容完全一致的检索记录了。

在“小数据量”的情况下，这种思维方式是可行的，毕竟供分析的数据还不多，那人们就需要尽可能地去精准分析和量化每一项记录。人们已经在某些方面意识到差别的存在了，比如晚上打烊时的小商店总是细细地把每分钱都数清楚，可是没有人会一分钱一分钱地去衡量GDP。随着数据规模的扩大，人们就不再会那么痴迷于精确度了。

专业的数据库是达到精确的基本保障。针对小数据和特定的某件事情，人们能够去追求精确度，像是某一个人想要确认自己的账户上是不是有足够的余额去开支票。可是大数据时代到来的时候，精确度的追求已经不可行，甚至不受到欢迎了。一旦有了海量的数据，人们所追求的目标就不再是精确度了。

复杂多样的大数据，在全球多个服务器上分布。人们在拥有大数据的情况下不用再纠结于一个特定的现象，只要是大方向掌握了就可以了。当然这不代表精确度就不需要了，只不过是不希望人们只是执着于此。微观层面上的精确度适当放弃一些，会让宏观层面上的我们有更好的洞察力。

前两个转变促成了第三个转变，那就是因果关系不再是人们热衷去寻找的了。人类长时间都习惯于寻找因果关系。即便是没有太多用途的因果关系，人们还是会习惯性地去找到这关系。可是到了大数据时代，

因果关系的重要性渐渐被忽略了，取而代之的是事物之间的相互关系，这显然是一个相当新颖且有价值的观点。相关关系或许有可能无法告知人们事物发生的原因是什么，但是能提醒大家有事情正在发生。很多情况下，这种提醒就足够了。

或许数以百万计的电子医疗记录说明特定的组合橙汁和阿司匹林对治疗癌症有作用，可是这种治疗方法显然要比找出具体的药理机制重要得多。相同的，我们只需要知道买机票的最佳时机是什么时候，完全没必要去理解价格疯狂变动的原因是什么。大数据能告诉我们“是什么”而不是“为什么”。现象背后的原因是什么不需要人们去理解，只要数据能够自我发声即可。

没必要用少量的假设来进行分析，如果海量收据还没有收集全的话。只要是数据能自我发声，那很多从前尚未为人们所意识到的联系就会被人们所认识的。

就好比对冲基金在分析了社交网站Twitter上的数据信息后推测了股市的表现，亚马逊和奈飞（Netflix）也在自己的网站上依据用户的类似查询来推荐商品，还有Twitter，Facebook和LinkedIn则是利用用户的社交网络来了解用户的喜好等等。

几千年前，人类就开始了数据分析。最初美索不达米亚平原上的记账人员发明了书写，只为有效跟踪记录信息。圣经时代开启后，为了人口普查，政府开始建立大型的国民数据库。两百多年来，精算师们也在大量收集数据，只为了规避风险。

在模拟数据时代，收集和分析数据是一件很困难的事情，一旦有新问题出现，数据分析就不得不重新开始。来到数据化的时代，数据管理效率已经有了大踏步的进步了。数据化的进程将原本模拟数据转化为了计算机可读的数据，存储和处理这些数据就不再那样困难了，数据管理的效率也逐步提高。曾经要几年才能完成的数据收集和分析，现在只消几天的工夫。不过改变不是唯一的。曾经人们有一种偏见，认为数据库仅有单一的用途和价值，那是模拟数据时代的技术和方法所导致的。大数据转变尽管是数字化促成的，但这绝不仅仅是计算机存在所能实现的。

现在正在发生的一切人们已经无法准确描述了，但“数据化”的概念足以让人们去了解这次变革。它意味着所有人都要从世间万物中汲取信息，包括那些曾经认定和“信息”完全没有关系的事情。例如人们所处的位置、引擎的震动、桥梁的承重等等。量化的方式能够让这些信息都化为数据，这也使得从前无法做到的事情如今都成了可能，例如引擎会不会出现故障可以从引擎的散热和震动来推测。这样一来，数据就能够挖掘出前所未有的潜在价值了。

大数据时代为人类提供了一场寻宝游戏，数据让人们将关注点从因果关系转为相关关系，并释放出了潜在的价值，这就是主宰这场游戏的关键。这一切成为可能都源于新兴技术工具的使用。宝藏中不止一件宝贝，数据集内部都有着不为人知的价值。这些数据价值的挖掘和利用的竞赛正在全球各地上演。

数据处理技术已不是新鲜的事物，但最初它只是调查局、研究所和世界上的巨头公司掌握的一项技术。像沃尔玛和美国第一资本银行就领先其他企业在零售业和银行业率先使用大数据，这两个行业因此而得到改变。现在这些技术已经普及到各个行业。

个人受到大数据的影响是最大的。专业性在一个可能性和相关性占主导的世界里已经不再重要了。不过这并不说明行业专家会消失，因为他们要和数据表达进行博弈，就像是《点球成金》电影里的那样，统计学家让棒球球探们感到惭愧不已，精准的数据分析让直觉的判断不得不退居二线。人们因此不得不去调整管理、决策、人力资源和教育方面的传统理念。

大部分时候，人们都会将习惯建立在一个预设好的立场上，但是这只适用于少量、精确且重要的决策信息时代。一旦数据量规模变大，处理速度增快且数据不再那样精确，所谓的预设立场就不存在了。数据量庞大的情况下，人类自身就要让步于机器来做出决策。当然，不得不承认这也存在一定的弊端。

社会在监视和了解人类的行为上已然有了数千年的历史了。可是一个算法系统要如何监测呢？信息化时代的最早期，政策专家已经发现了人类的隐私权遭受到了信息化的威胁，人们不得不建立起一个庞大的规则体系来维护个人的信息安全。大数据时代来临的时候，这些规则又都失去了原有的作用。对于信息分享人们很热衷也都是自愿的，网络服务当中的一个中心特征就是信息分享，而不再是规避的薄弱点了。

对人们来说，似乎隐私的泄露已经不是危险，而转换成了被预知的可能，它们能够预测人们是否拖欠还款或是犯罪，这些算法很显然会让人们在尚未做出这些行为之前就预先被逮捕。因此大数据是统计的首位，那即便是这样，个人意志是不是就高于大数据呢？就仿佛出版印刷行业的发展推动了一个国家言论自由的立法保护，而大数据时代也必须推动个人权利捍卫的立法保护才行。

在处理和控制数据的方法上，政府机构和社会都要有全方面的改变。这是一个难以否认的用数据来预测一切的年代，尽管背后的原因究竟是什么我们难以解释。医生若是要求病人遵从医嘱，但对于医学干预的合理性无法说明的话，那将是什么情况呢？事实上在大数据时代，这是懂得病理分析的医生一定会去做的事情。此外司法系统的“合理证据”要不要改成“可能证据”呢？若可以这样做的话，那么人类的自由和尊严会不会也因此受影响呢？

大数据时代也必须有一系列倡导的规范制度，它们和曾经小数据时代发展并留存下来的规范有着继承的关系，只不过在新环境下旧规范也要与时俱进了。

人类在寻求量化方式认识世界的道路上大跨步迈进的标志就是大数据的出现。从前很多东西无法计量、存储、分析和共享，如今都被数据化了。世界因为大数据和海量的不如从前精确的数据为人们开启了一扇全新的理解世界的大门。社会对于因果关系不再执着，反而去关注相关关系的好处。

现代社会有了寻找原因这一神论，但因大数据的出现而被推翻。只不过人们又陷入了另一个历史困境中，那便是这是一个“上帝已死”的时代。换句话说，传统固守的信念已经开始发生动摇了。更让人感觉到讽刺的是，信念正在为“更好”的证据所取代。经验中获得的信息、直觉和迷惘都和证据相矛盾，它们是什么样的角色呢？世界转而开始挖掘相关信息，人们又该如何在不破坏传统因果推理的基础上来保证社会繁荣、人类前进并取得实质性的进步呢？